Development of an oscillation detection method and establishment of enzyme characterization system using mass spectrometry data

Hitoshi Iuchi

Graduate School of Media and Governance,

Keio University

This dissertation is submitted for the degree of

Doctor of Philosophy

Abstract

The big issue of systems biology is how to extract biological significance in *omics* datasets. In this thesis, I describe an algorithm to detect oscillating molecules in *omics* datasets for the study of circadian rhythms, and enzyme characterization of the orphan enzyme YhhY in *Escherichia coli* (*E. coli*) by mass spectrometry. In Chapter 1, I present a historical overview of research on circadian rhythms and show that it is essential to adopt a systematic approach in the research on circadian rhythms. In Chapter 2, I describe maximal information coefficient-based oscillation prediction (MICOP), which is an oscillation prediction method using mutual information coefficient. Owing to the development of parallel sequencers and mass spectrometers, classical methods are not suitable for *omics* datasets that have a low sampling rate and contain noise. From the results of numerical experiments, I show that MICOP demonstrates excellent performance for data with a low sampling frequency and noise, such as time-series *omics* data. Furthermore, upon the application of MICOP to a plurality of published mouse proteome data, novel oscillating molecules were identified, with the results being shown to be roughly in agreement with the transcriptome results. From these results, I assert that MICOP is an excellent oscillating molecule detection method. In Chapter 3, I describe a functional identification of the *E. coli* function-unknown gene *yhhY*. Although YhhY was predicted to be an acetyltransferase from its amino acid sequence, identification of the substrate and the obtainment of experimental proof of its function have not been conducted. Therefore, in this study, I aimed to clarify the function of YhhY by combining biochemical experiments and metabolome analysis using a mass spectrometer.

The results suggested that YhhY is an enzyme that specifically acetylates some essential amino acids and that it is involved in metal stress *in vivo*. Finally, in Chapter 4, I summarize the contents of this paper.

Keywords: circadian rhythms, time-series data, mass spectrometry, YhhY

要旨

システムズバイオロジーにおいて、オミクスデータから生物学的知見をいかにして得るかは重要な課題である。本論文の前半部分では概日リズム研究においてオミクスデータから振動している分子を検出するためのデータ解析技術について、後半では質量分析計を用いた大腸菌機能未知酵素 YhhY の機能同定について記述する。第1章では、概日リズム研究の歴史を俯瞰し、概日リズムが複数のレイヤーをまたいだ分子のネットワークから構成されていることが明らかになってきたこと、それに伴い概日リズム研究にシステムバイオロジー的アプローチが必要になってきたことについて述べる。第2章では相互情報量を用いた振動推定法である Maximal information coefficient-based oscillation prediction (MICOP) について述べる。並列シーケンサーや質量分析計の発展によって、従来の統計手法では対応できない事例が増えている。それに対して、MICOP は時系列オミクスデータのような、サンプリング頻度が低くノイズが多いデータに対して優れたパフォーマンスを発揮することを数値実験の結果から示した。さらに、公開されている複数のマウスプロテオームデータに MICOP を適用したところ、新規振動分子が同定され、かつそれらの結果はトランスクリプトームの結果と概ね一致していることが分かった。これらの結果から、MICOP は優れた振動推定法であることを主張する。第3章では大腸菌機能未知酵素 YhhY の機能同定について述べる。YhhY はそのアミノ酸配列からアセチル基転移酵素であることが予測されていたが、基質の同定やその機能の実験的証明は行われていない。本研究では、生化学実験と質量分析計によるメタボローム解析を併用することで、YhhYの機能を明らかにすることを目指した。その結果、YhhY はいくつかのアミノ酸を特異的にアセチル化する酵素であること、さらに生体内においては金属ストレスに関与していることが示唆された。第4章では、本論文の内容を総括する。

キーワード: 概日リズム、時系列データ、質量分析計、YhhY

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ARS Autoregressive spectral analysis

CE-MS Capillary Electrophoresis-Mass Spectrometry

DD Dark-dark cycle

FDR False discovery rate

FN False negative

FP False positive

JTK Jonckheere-Terpstrz-Kendall algorithm

LC-MS Liquid Chromatography-Mass spectrometry

LD Light-dark cycle

LS Lomb-Scargle

MCC Matthews correlation coefficient

MIC Maximal information coefficient

MICOP Maximal information coefficient-based oscillation prediction

PCA Principle component analysis

ROC Receiver Operating Characteristic

SCN Suprachiasmatic nucleus

SDS-PAGE Sodium dodecyl sulfate-Polyacrylamide gel electrophoresis

TN True negative

TP True positive

Chapter 1

General introduction

1.1 What is systems biology

Large-scale experiments such as genome sequencing or expression profiling using a parallel sequencer and functional screening by mass spectrometry have produced huge data. The growing demands for interpretation of these datasets and for their integration to obtain a deeper understanding of life have created the systems biology. The big issue of systems biology is how to connect large-scale data to a comprehensive understanding of life. The mass spectrometry or parallel sequencer output complex data, but there are still many problems to make efficient use of the data. Thus, how to extract biological significance in large datasets measured by mass spectrometry or parallel sequencer is a big issue of systems biology. In this thesis, I describe the development of an oscillation detection method and the establishment of enzyme characterization system using mass spectrometry data.

1.2 Circadian rhythms and systems biology

1.2.1 Historical overview of circadian rhythms

The ability to adapt to the light-dark cycle on the Earth provided a survival advantage to many organisms. These endogenous 24 hour rhythms were named circadian ("about" and "a day") rhythms [1]. The first arrhythmic mutants were found in *Drosophila melanogaster* by ethyl methanesulfonate-induced mutant screening [2]. Some groups identified that *period* gene, which would be the best known later, was responsible for their arrhythmic behavior at almost the same time [3–5]. In mammals, the destruction of the suprachiasmatic nucleus (SCN) located in the hypothalamus was also shown to abolish behavioral and hormonal rhythms, whereas the rescue of SCN led to a recovery of circadian rhythms; these findings revealed that SCN is the master clock in mammals [6, 7]. Moreover, the screening of mutant mice using *N*-ethyl-*N*-nitrosourea revealed arrhythmic mutants, followed by identification of the first clock gene *circadian locomotor output cycle kaput (Clock)* in mammals [8–10].

1.2.2 Systems biology and circadian rhythms

Mammalian circadian rhythms are good examples of systems biology because it is a typical model system which is complex and dynamic. The circadian clock was defined using the following three general criteria [11].

The rhythms are endogenous and oscillatory with an approximately 24-hour period.

The rhythms persist under constant conditions such as constant darkness with an approximately 24-hour period.

The rhythms are entrainable to external environmental changes.

The rhythms are reset by environmental stimulation such as light and dark. This criterion confers the ability to adapt to the environment and jet lag, for example.

The rhythms are temperature-compensatory.

Many organisms live at a broad range of temperatures, but the circadian rhythms are always approximately 24 hours. The mechanism behind this remained unclear, and many issues about this were still unresolved. However, recent work has described that this system can be explained by enzyme kinetics, without the need for transcription and translation mechanisms to be involved [12].

The characteristics mentioned above are achieved by multilayered mechanisms, which are difficult to clarify without systematic approaches (Fig. 1.1).

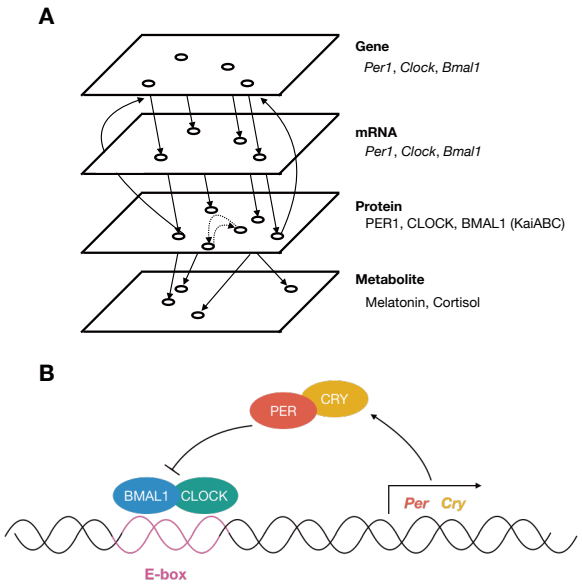

Fig. 1.1 Multilayered mechanisms to construct circadian rhythms
(A) In mammals, the transcriptional/translational feedback loop is a core part of the architecture of circadian rhythms. (B) Typical regulation of circadian rhythms by the transcriptional/translational negative feedback loop.

1.2.3 Molecular timetable to predict internal body clock

Circadian rhythms control many physiological functions and metabolic processes [13]. For example, hormone secretion is under the control of the circadian rhythms. The abundances of melatonin and cortisol in the blood, for example, oscillate in a circadian manner to regulate the daily functions of peripheral organs [14–18]. There is increasing evidence that dosing time affects drug efficacy and toxicity in humans [19–24]. For example, appropriately timed administration of two anticancer drugs in ovarian cancer patients (adriamycin at 6:00 AM and cisplatin at 6:00 PM) resulted in lower toxicity than arrhythmic administration [25]. However, it has been reported that individual internal body times can vary by up to 5–6 hours among healthy humans and even by 10–12 hours for night-shift workers [26–29]. The conventional method to estimate body time is to measure the blood level of melatonin or cortisol over 24 hours because of the abundance of these hormones fluctuates in a circadian manner [14, 17, 18]. However, this requires periodic blood sampling in a controlled environment, which severely restricts patients' activity. A simpler method is needed to measure patients' body time in order to optimize drug efficacy [19–21, 23, 30]. Ueda *et al.* previously demonstrated that a molecular metabolite timetable inspired by Linnaeus' flower clock could resolve these issues [14, 31]. In Linnaeus' flower clock, the pattern of opening or closing of different flowers indicated the time of day. Likewise, they constructed a timetable of physiological molecules (e.g., metabolites) at particular times of the day and showed that this timetable could accurately estimate body time.

For constructing the timetable, they used metabolome techniques to profile hundreds to thousands of molecules in a sample. Metabolome profiling by mass spectrometry has previously been used in other research fields including studies of bacteria, plants, and mammals [32, 33]. The recent integration of capillary electrophoresis and mass spectrometry (CE-MS) has achieved a high resolution with short measurement time [32, 34]. CE-MS is capable of identifying several hundred metabolites in a small sample volume (on the order

of nL). However, CE-MS is not suited to the measurement of neutral metabolites such as fatty acids and sugars because the technique is optimized for ionic molecules. Instead, liquid chromatography-MS (LC-MS) fills the gaps left by CE-MS [32].

1.3 Statistics to detect rhythms in chronobiology

A major issue in the field of circadian rhythms is how to detect circadian periodicity, for which the χ square periodogram is the golden standard for circadian rhythm studies [35, 36]. This approach has been widely used for circadian cycle analysis of behavior, such as a locomotor activity of mouse and rat or electric activity of cultured neurons/tissues. The χ square periodogram was applied for time-series data using the following procedures (Fig. 1.2).

1. Detrending by removing the head of the time-series data
2. Separating the time-series data by assumed periods P
3. Calculating the mean data $\overline{y_h}$ at each time point by superimposing each section
4. Testing the periodicity by calculating Q_P, which was the normalized square error of $\overline{y_h}$ and $\overline{y}$, and comparing with the significance level

Q_P is calculated using the following formula,

$$Q_P = \frac{KN \sum_{h=1}^{P} (\overline{y}_h - \overline{y})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} \tag{1.1}$$

where P is the period of samples, $\overline{y}_h$ are the means of column after arranging the series (of n elements) in an array of P columns, and K is the row number of the resulting array. Q_P follows a χ^2 distribution with as many degrees as cycles in each section.

Q_P becomes smaller when the assumed period P is out of the period of the actual time-series, and it is maximized when it coincides with each other. It is expected that Q_P will follow the χ^2 distribution whose degree of freedom is $P-1$, if the time-series data are long, and it is assumed that each point is independent of the same mother distribution. The common procedure is that Q_P is calculated when changing the assumed period P (14–34 hours in this study), and Q_P is plotted with statistically significant levels (Fig. 1.2).

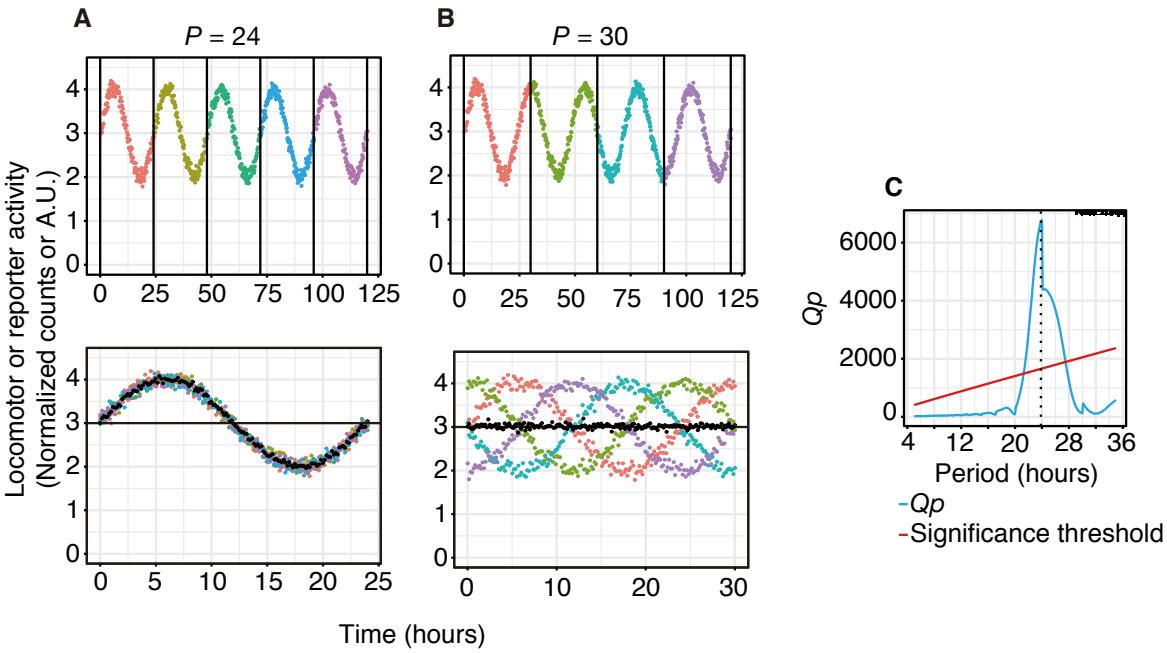

Fig. 1.2 Procedure of χ^2 periodogram
(A) Time-series data were separated with an assumed period $P = 24$. (B) Time-series data were separated with an assumed period $P = 30$, and they were overlaid. (C) Q_P was calculated with gradual change of the assumed periods. The red line indicates the significance level.

Chapter 2

MICOP: Maximal information coefficient-based oscillation prediction to detect biological rhythms

2.1 Background

The circadian rhythm, which involves oscillations over a cycle lasting 24 hours, plays a critical role in biological systems [37]. Transcriptional negative feedback loops composed of clock genes are a key component of this mechanism [37–39]. These clock genes regulate downstream gene expression, leading to the 24-hour rhythms of various physiological phenomena such as cell division, energy metabolism, blood pressure, and sleep [17, 18]. Many molecules are involved in these systems, so comprehensive and multilayered approaches are required to clarify the complex systems. Thus, it is crucial to obtain a deep understanding of the circadian rhythms in order to understand biological systems.

The availability of biological time-course data is key to elucidating circadian rhythms, but there are several difficulties in analyzing biological time-series data. In particular, the accumulation of time-series *omics* data via the technological innovation of mass spectrometry

and DNA sequencers has led to the following problems: (1) low sampling frequency and (2) unstable oscillation. The first problem is derived from the generally low sampling frequency of *omics* datasets because comprehensive approaches such as proteomics and transcriptomics are often expensive and laborious. Several *omics* studies collected time-course data every 2–4 hours per day and estimated periodicity using 12 to 24 points [14, 31, 40, 41]. This sampling frequency of *omics* data was relatively low compared with those for locomotor activity or tissue luminescence, which were provided every minute [42]. The second problem is the amplitude decay of time-course experimental values. There are various types of unstable oscillations in the expression patterns of genes and proteins. For example, previous reports assumed unstable oscillations such as cosine with outlier time points, cosine with a linear trend, cosine with an exponential trend, and decaying cosine as possible natural oscillation phenomena [42, 43]. These unstable oscillations hamper oscillation detection, in particular for amplitude decay, which is often observed in experimental systems, and is caused by degradation of the metabolic activity of cells and degradation of fluorescent proteins [44]. Therefore, the novel computational method that functions over the time course of *omics* studies with limited sampling points and amplitude decay should be developed.

Many analytical approaches to predict molecules with oscillating levels from time-series data have been developed. These algorithms were classified into time-domain and frequency-domain methods [12]. Typical time-domain methods are based on cosine curve-based pattern matching, and their simple algorithms help biologists to evaluate their analytical results [45]. For example, chi-squared periodogram is algorithms employing curve fitting and autocorrelation, respectively [46]. Hughes *et al.* developed Jonckheere-Terpstrz-Kendall algorithm (JTK) which is a nonparametric approach using rank by the nonparametric Jonckheere-Terpstra test and obtained the strength of correlation by Kendall's tau test [47]. However, they have disadvantages, such as sensitivity to noise and outliers, and being able to detect only cosine wave-like curves; as such, there is a need for a

novel algorithm that can overcome these obstacles. Meanwhile, frequency-domain methods based on spectral analysis are strongly noise-resistant and model-independent; however, the results are difficult for biologists to interpret [45]. Fisher's *G*-test estimates periodicity by calculating the periodogram of experimental data and calculating the *P*-value using Fisher's *G*-statistic [48]. Autoregressive spectral analysis (ARS) is an approach combining time-domain and frequency-domain methods, used to identify molecules with rhythmically oscillating levels in large-scale time-resolved profiles by autoregressive spectral analyses [49, 50]. Similarly, an approach combining autocorrelation and spectral analysis after removing noise from raw data with a digital filter was also proposed [51]; however, frequency-domain methods are limited by the low sampling frequency and short period in *omics* experiments, which means that they are often insufficient to predict the periodicity of large-scale *omics* datasets [52]. Therefore, developed approaches to characterize oscillating molecules in biological data have been used with success and have contributed to our understanding of biological systems; meanwhile, it has been shown that each method sometimes produces inconsistent results because of noise, sampling rate, and waveform [53]. A novel oscillation prediction method compatible with *omics* experiments, having a low sampling frequency, was required, for which a quantitative evaluation of the performance could also be achieved.

This study developed the maximal information coefficient (MIC)-based oscillating prediction (MICOP) for analyzing time-series *omics* datasets with high-level noise and possible decay. MICOP offers unsurpassed performance to identify and characterize oscillating molecules in *omics* datasets.

2.2 Materials and methods

2.2.1 Datasets

Time-resolved data from biological samples are generally obtained every 2–6 hours per day [14, 31, 40, 41]. Therefore, I simulated time-series data containing 6–24 points for two cycles for a performance test. Half of these artificially simulated data did not feature oscillation, while the other half did. For oscillating data, to mimic experimental data, noise according to the normal distribution (average = 0, the standard deviations = 0–0.6) was added to the sine curve. The decaying time-series datasets were designed so that the value of the peak in the second cycle is one-third of the value of the peak in the first cycle. The non-oscillating data were random numerical data. Proteomics datasets of C57BL/6J and C57BL/6 were downloaded from journal websites [40, 41]. The simulated data released by Wu *et al.* are included in MetaCycle, as described below [53, 54].

2.2.2 Design

A conceptual diagram of MICOP is shown in Fig. 2.1.

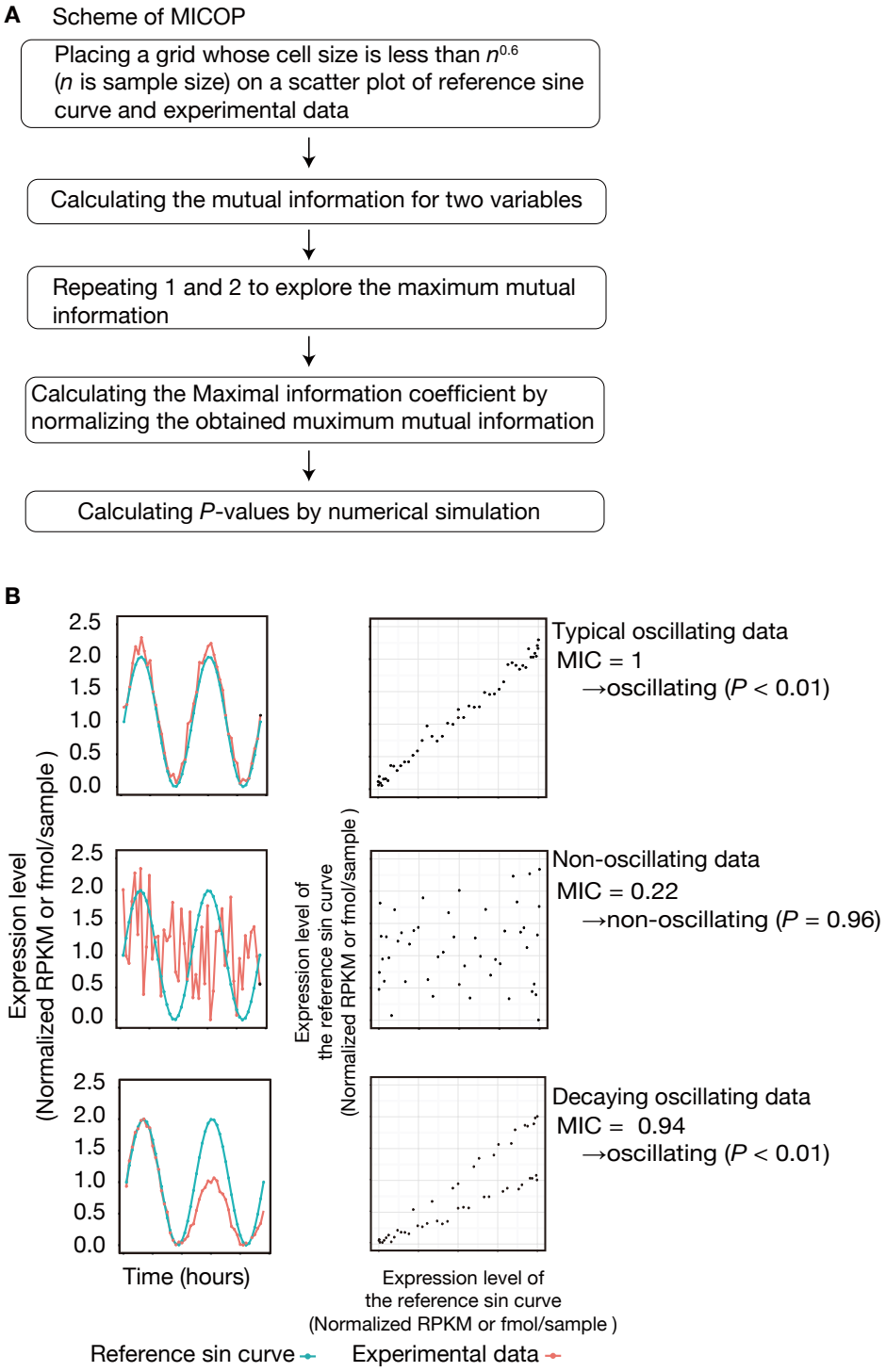

Fig. 2.1 A conceptual diagram of MICOP

(A) Procedure of MICOP. (B) Left boxes: experimental data (red) and reference sine curves (blue); right boxes: scatter plots between reference sine curve (x-axis) and experimental data (y-axis); top: typical oscillating data (MIC = 0.1, $P < 0.05$); middle: non-oscillating data (MIC = 0.22, $P > 0.05$); bottom: decaying oscillating data (MIC = 0.94, $P < 0.05$).

The MIC belongs to the nonparametric exploration class, and the score indicates the strength of the linear or non-linear association between variables. First, the mutual information for a scatter plot of X and Y is calculated as:

$$I(X;Y) = \sum_{Y}\sum_{X} p(X,Y)\log_2 \frac{p(X,Y)}{p(X)p(Y)} \tag{2.1}$$

where *p(X)* and *p(Y)* are marginal probability distribution functions of X and Y, and *p(X,Y)* is joint probability distribution function. Then, to compare the values from different grids and to obtain normalized values between 0 and 1, MIC is divided by the lesser number of X and Y bins. MIC is calculated as:

$$MIC(X,Y) = \max_{X,Y<n^{\alpha}} \frac{I(X;Y)}{\log_2(\min(X,Y))} \tag{2.2}$$

The algorithm calculates the MIC value between the reference sine curve and experimental data. The same sine curve was used for all input traces. The script for MICOP and its performance test is provided as an R script. The *P*-values were calculated from the frequency of each MIC value of experimental data and the MIC values that were calculated from the random numbers. The MIC represents the strength of the association between the two variables. The MIC between the reference sine curve and targeted data, such as experimental data or simulated data, was calculated using the following steps. Step 1: Grids with different resolutions are introduced to separate the different areas of the scatter plot of the two variables. Step 2: Maximized mutual information at each resolution is selected. Step 3: The mutual information is normalized for each resolution. Step 4: The maximum value among all division methods is MIC. Step 5: to calculate the *P*-value, MIC between the reference curve and 1000 non-oscillating time-series datasets, which comprised random values, was calculated. I compared MIC values and enumerated the occurrences (k) when the MIC score exceeded the score calculated. k/1000 was taken as the *P*-value of the MICOP. Then, I

compute the *P*-value as:

$$P = \frac{1}{1000} \sum_{t=1}^{1000} I(MIC(Xpi, Ypi) > MIC(X, Y)) \tag{2.3}$$

where I is the indicator function, and Xpi and Ypi is the ith permutated version of X and Y, respectively. If the datasets have missing points, MIC is calculated without the point. The *P*-values were corrected by the Benjamini–Hochberg method [55].

2.2.3 Performance test

The periodicity of the simulated data was determined by MICOP, JTK, ARS, and Lomb-Scargle (LS) to test the performance of MICOP. The Matthews correlation coefficient (MCC) was compared to compare the precision and sensitivity of MICOP. MCC values were calculated as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.4}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, *FN* is the number of false negatives. The false discovery rate (FDR) is widely used and is calculated from true positive and false positive values. In contrast, MCC is more informative as a value evaluating the performance of the classification method because it is calculated from the number of *TP*, *FP*, *TN*, and *FN*.

2.2.4 Reanalysis of previously reported time-resolved proteomics datasets

To verify the practicality of MICOP, I reanalyzed the published time-series data [40, 41, 54, 56, 57]. Briefly, these are proteome datasets of C57BL/6, or C57BL/6J liver sampled

every 3 hours for 1–2 days. Except for Robles *et al.*, 2017, mice were bred under the light-dark condition, and mouse in Robles *et al.*, 2017 were bred under constant dark condition. Moreover, simulated data which are two cycles containing 20 molecules. The MIC and *P*-value were calculated as described in the Design section.

2.2.5 Programming languages and statistical analysis

R language (ver. 3.3.2) was used for all analyses [58]. Three different random seeds were used; rnorm function was used to generate random numbers according to a normal distribution, and runif function was used to generate uniform random numbers. The performance of MICOP was compared to each method by Welch's *t*-test. The Benjamini–Hochberg procedure corrected the *P*-values for multiple testing. A graphical package named ggplot2 (ver. 2.2.0) was used to draw figures. The Minerva package (ver. 1.4.3) was used to calculate the MIC score, and binning range to calculate MIC score was 0.6, which is a default value of the R library. The MetaCycle package (1.1.0) was used for periodicity judgment by ARS, JTK and LS [51, 53, 54, 59, 60].

2.3 Results

2.3.1 Parameter search for MICOP

First, I tested gradually changed α to optimize MICOP for time-series analysis (Fig. 2.2). As a result, MCC values were almost constant without α, indicated that α is not essential for oscillation prediction. In this paper, α=0.6 which was recommended by the original article of the MIC was used for the following analysis.

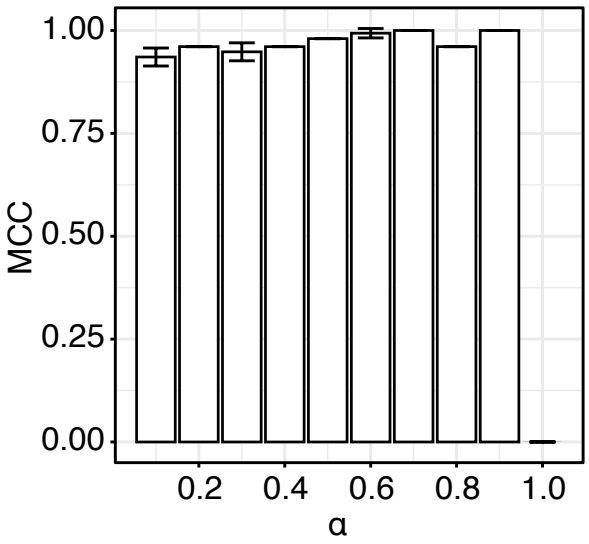

Fig. 2.2 Optimization of α for time-series analysis
The x-axis indicates α and the y-axis indicates MCC values. MCC values and α have no units. The error bars indicate standard error (n=3).

2.3.2 Computation time of MICOP

To evaluate computation time for evaluation of time-series data, I compared CPU time of MICOP and widely used existing methods (Fig. 2.3). The process times of all methods were sufficiently shorter than experiment containing mice sampling and proteome analysis by mass spectrometry. The characteristic point of MICOP was that the calculation cost increased exponentially as the sampling point increased.

2.3.3 Noise levels of proteome data

To evaluate the actual noise level of proteome datasets, I calculated noise level of published mouse proteome data (Fig. 2.4) [61]. The average noise levels of detected proteins were approximately 0.2, and the noise levels of typical clock proteins were between 0.1–0.3.

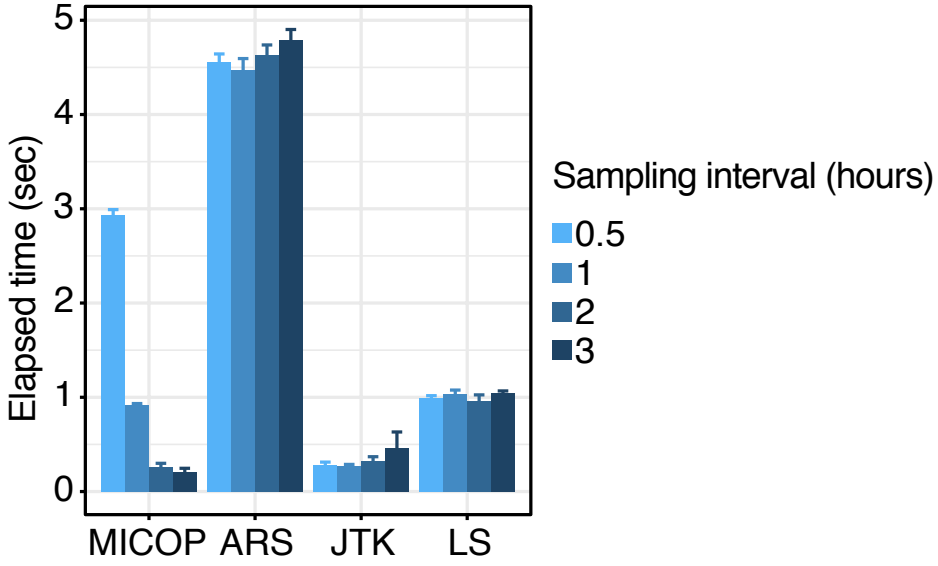

Fig. 2.3 Comparison of computation time of MICOP and existing methods for two-cycle data

The x-axis indicates each method and the y-axis indicates elapsed time (sec). Sampling intervals are illustrated as color gradation. The error bars indicate the standard deviations (n=3).

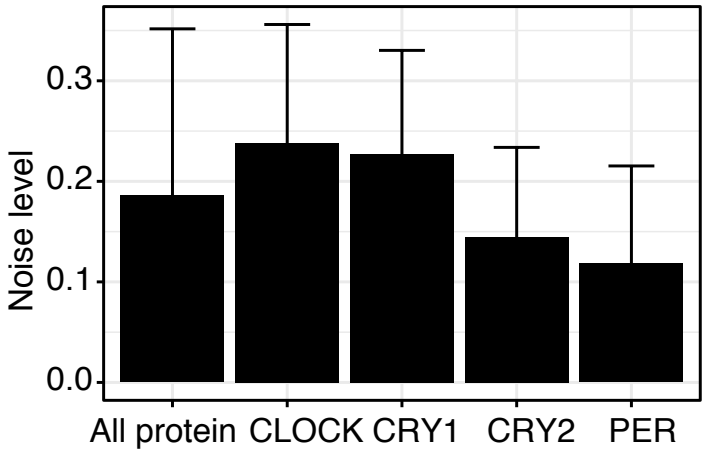

Fig. 2.4 Noise levels of proteome data
Noise levels of all protein and typical clock proteins were shown. The y-axis indicates noise level. Noise level has no units. Error bars indicate the standard deviations (n=3).

2.3.4 Comparison of MICOP and existing methods for decaying data

To test the performance of MICOP, JTK, ARS, and LS for mimicking the decaying time-resolved data, the MCC values were calculated to differentiate significantly oscillating data from non-oscillating data using time-series simulation data, including 100 sets of oscillating data and 100 sets of non-oscillating ones (Figs. 2.5 and 2.6). Two-way ANOVA with method and sampling frequency as factors revealed significant effects of method ($F = 18.14$, $P < 8.28 \times 10^{-8}$), sampling frequency ($F = 47.53$, $P < 1.64 \times 10^{-8}$) and method $\times$ sampling frequency interaction ($F = 14.55$, $P < 1.01 \times 10^{-6}$). MCC values were 0.72 ($P < 0.005$), 0.40 ($P < 0.005$), 0.082 ($P < 0.005$), and 0.00 ($P < 0.005$) for MICOP, ARS, JTK, and LS, respectively, when the sampling interval was 4 hours (Fig. 2.5). The MCC values increased as the sampling frequency increased, and these values became almost equal to 1 in all methods at 1-hour interval sampling. The MCC values of MICOP were 0.7 or more at all sampling frequencies and were the highest at sampling intervals of 1–3 hours, followed by ARS and JTK. LS did not function as a classifier at sampling intervals of 1–3 hours. Furthermore, Receiver Operating Characteristic (ROC) curves were consistent with these results (Fig. 2.7).

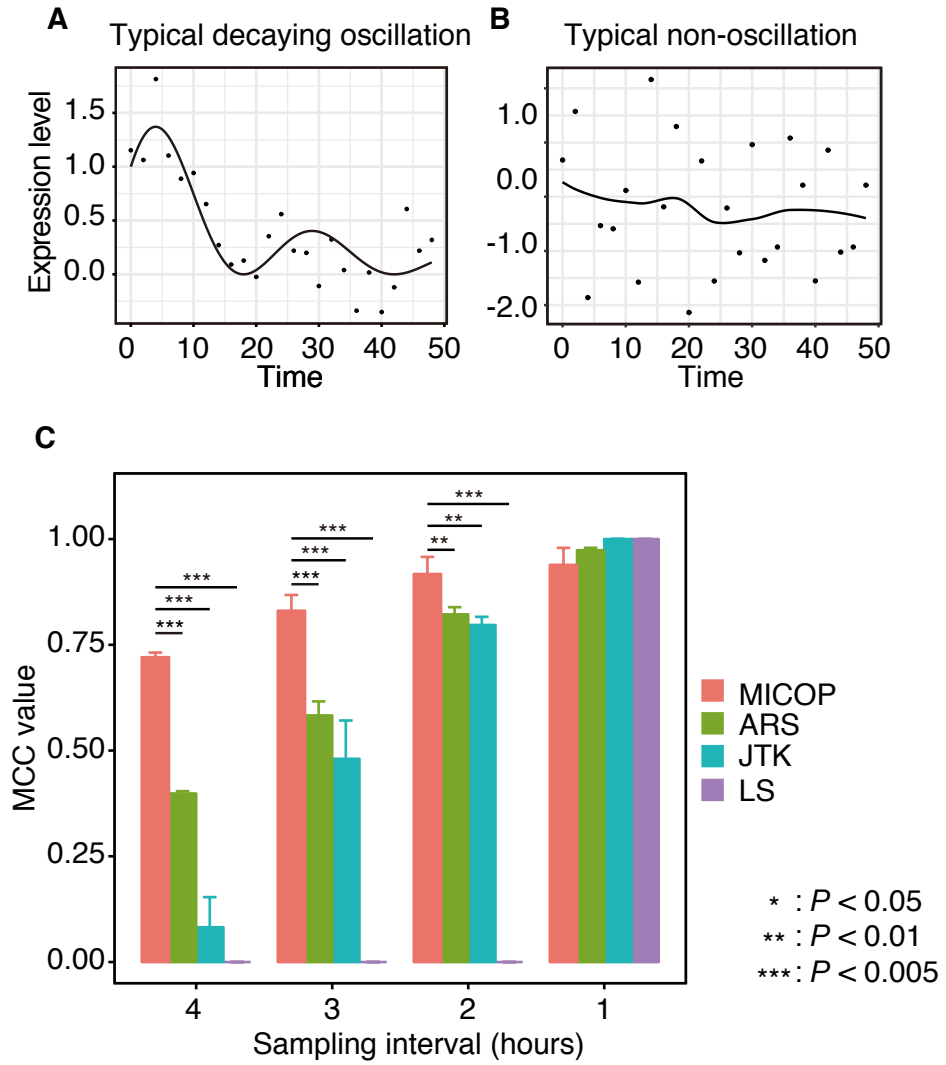

Fig. 2.5 MCC values of MICOP, ARS, JTK, and LS for decaying data
Comparison of detection power of MICOP and existing methods for decaying data. (A) Typical decaying oscillation data, (B) typical non-oscillation data, (C) MCC values from simulated time-resolved data in which half represent oscillating data, whereas the other half represent random numerical data, of which half does not oscillate. Noise level was 0.4 (the standard deviation). The x-axis represents the MCC value, while the y-axis represents the sampling interval (hours). The color indicates each method: red, MICOP; green, ARS; blue, JTK; and purple, LS. Error bars indicate the standard deviations (n=3). MCC and noise level have no units.

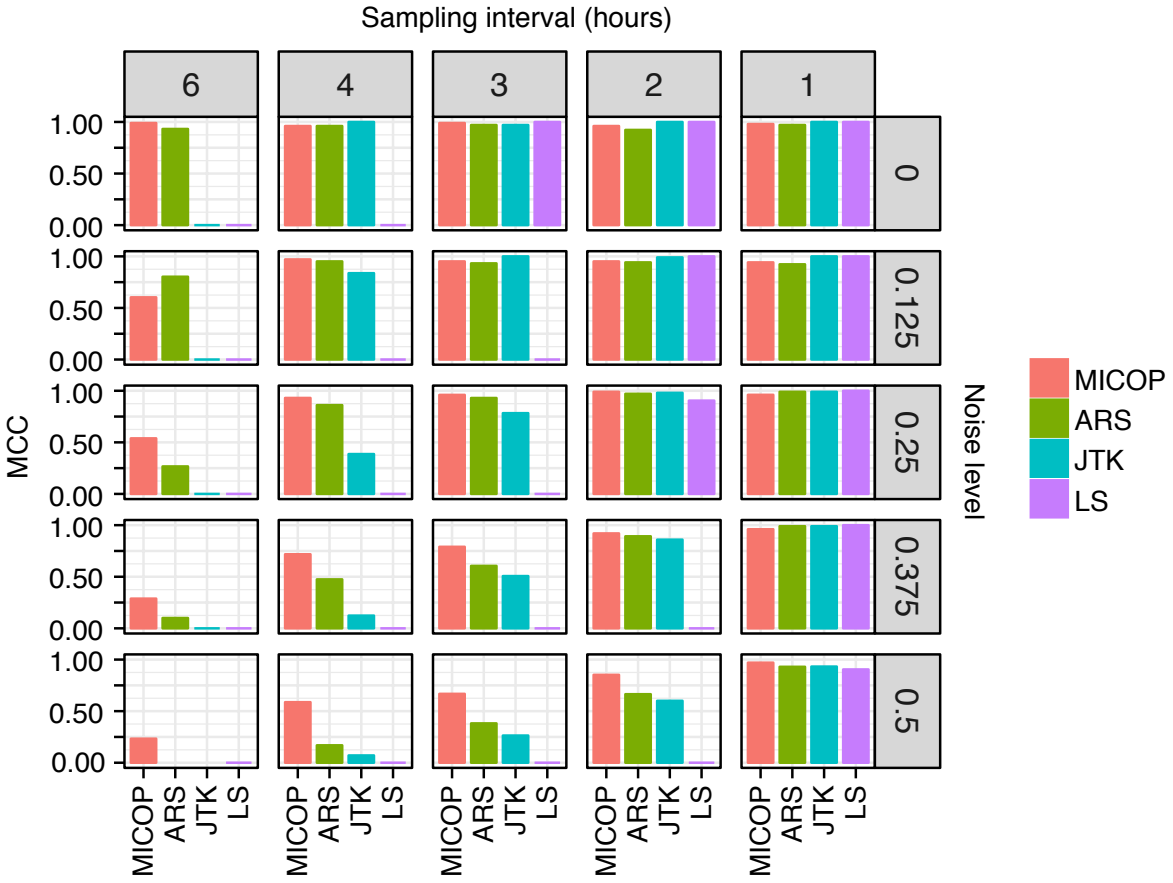

Fig. 2.6 Wide range comparison of MCC values of MICOP, ARS, JTK, and LS for decaying data

Sampling intervals and noise levels were gradually adjusted. The bar indicates MCC values (1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a prediction in complete disagreement). MCC and noise level have no units.

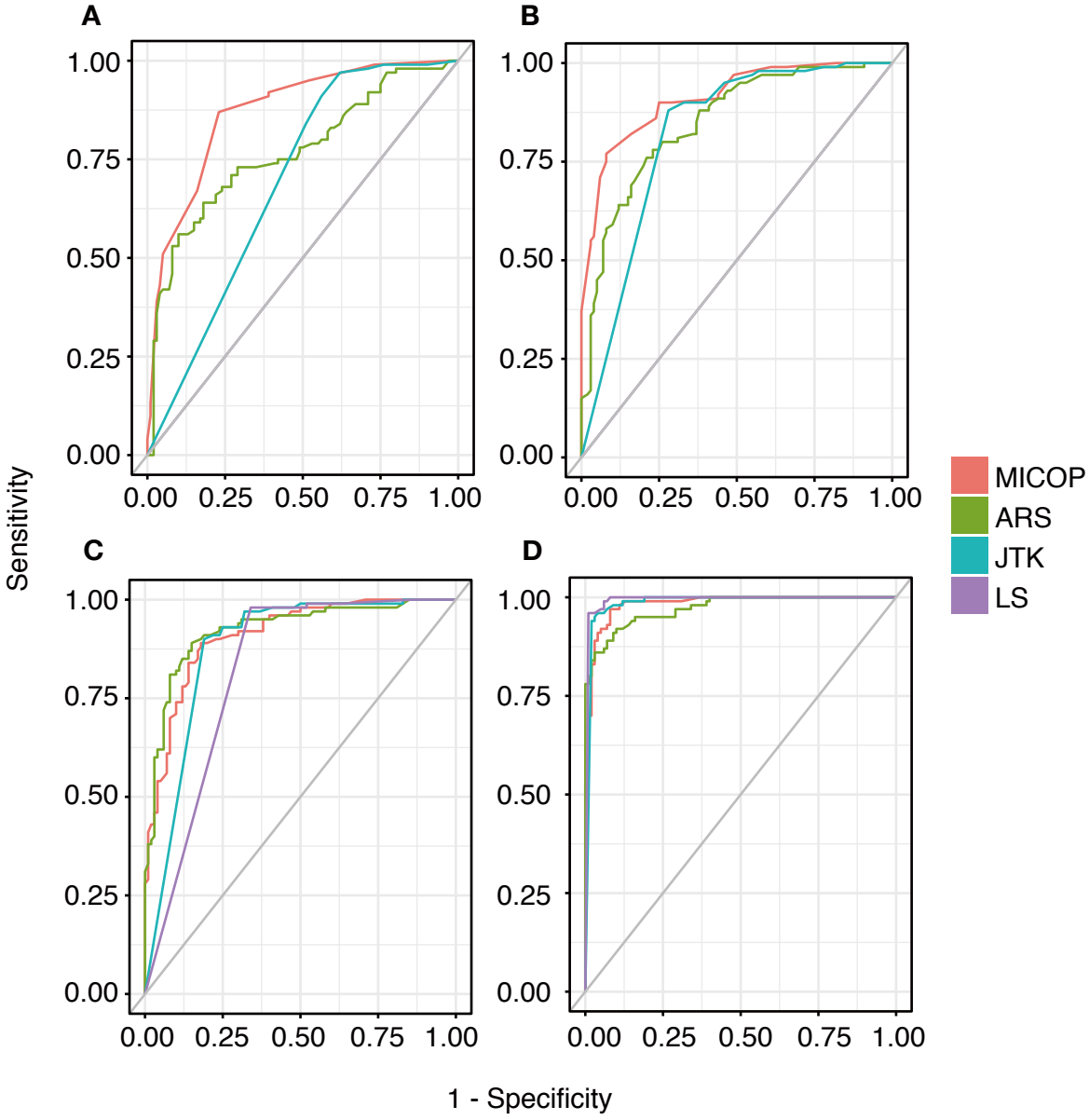

Fig. 2.7 ROC curves for time-series data with different sampling frequencies or gradually added noise

ROC curves of MICOP, ARS, JTK, and LS for decaying data ROC curves from simulated time-resolved data in which half represent oscillating data, whereas the other half represent random numerical data, of which half does not oscillate. The color indicates each method: red, MICOP; green, ARS; blue, JTK; and purple, LS. Each panel has different sampling intervals, sampling intervals are (A) 4 hours, (B) 3 hours, (C) 2 hours and (D) 1 hour. Sensitivity and specificity have no unit.

2.3.5 Comparison of MICOP and existing methods for noisy or low-sampling-frequency

I compared the accuracy of MICOP and existing methods for time-series data containing noise and having a low sampling frequency without attenuation (Fig. 2.8). Initially, I quantitatively evaluated the degradation of classification performance due to the noise of MICOP (Fig. 2.8). Two-way ANOVA with method and noise level as factors revealed significant effects of method ($F = 1.09 \times 10^3$, $P < 2.00 \times 10^{-16}$), noise level ($F = 6.43 \times 10^2$, $P < 2.00 \times 10^{-16}$) and method $\times$ noise level interaction ($F = 4.76 \times 10^2$, $P < 2.00 \times 10^{-16}$). The MCC values were 0.8 or more, except for LS, in all conditions, even if the noise was 0.500; however, LS did not function as a classifier when the noise was 0.375 or more.

The performance of MICOP as a classifier for low-sampling-frequency unattenuated data was also quantitatively evaluated (Fig.2.8). Two-way ANOVA with method and sampling frequency as factors revealed significant effects of method ($F = 4.24 \times 10^2$, $P < 2.00 \times 10^{-16}$), sampling frequency ($F = 4.48 \times 10^2$, $P < 2.00 \times 10^{-16}$) and method $\times$ sampling frequency interaction ($F = 1.42 \times 10^2$, $P < 2.00 \times 10^{-16}$). The MCC values increased in all methods as the sampling interval decreased, and were equal to 1 in all four methods at a sampling interval of 1 hour. LS did not function as a classifier at sampling intervals of 3–4 hours. The MCC values of MICOP were 0.7 or more under all conditions.

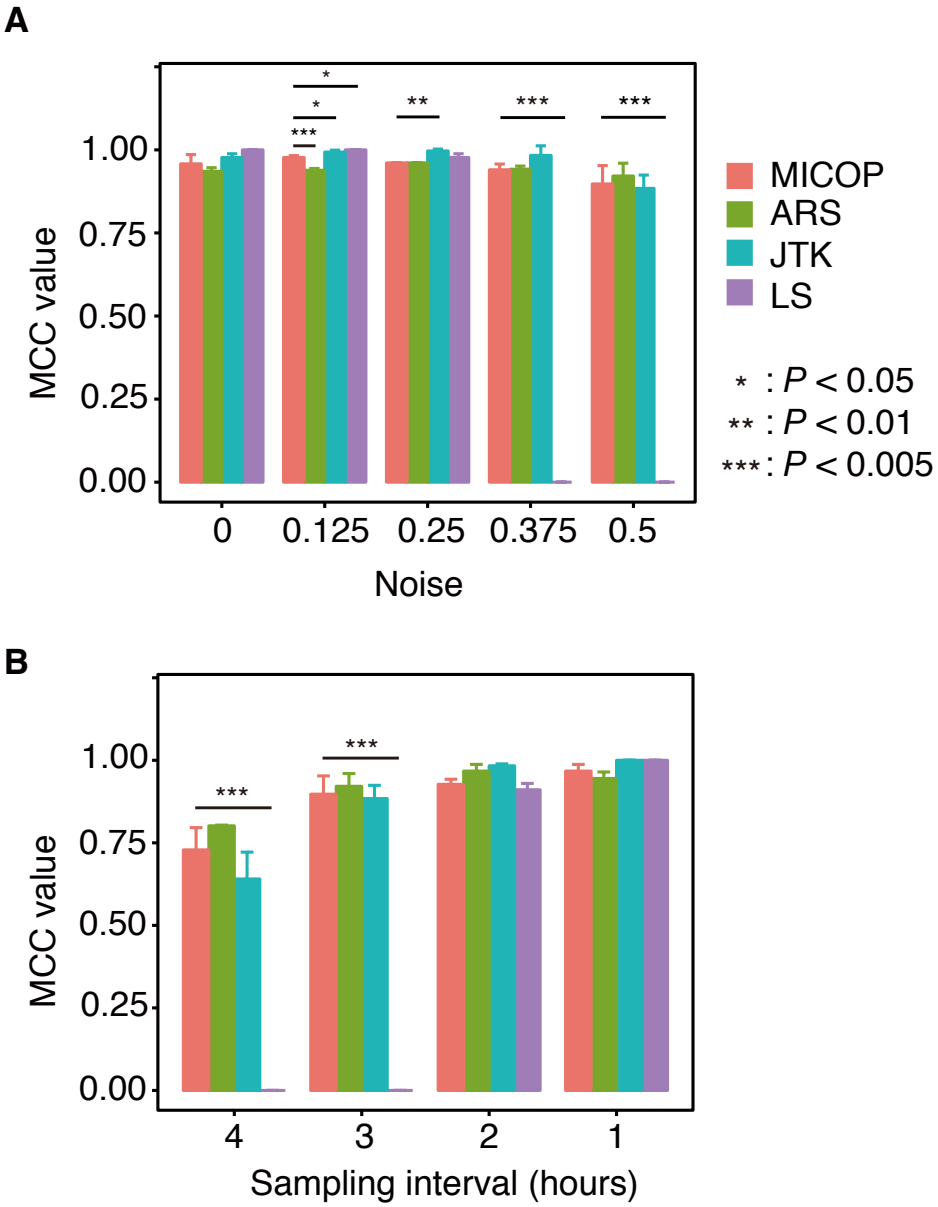

Fig. 2.8 MCC values for time-series data with different sampling frequencies or gradually added noise without attenuation

Comparison of MCC values of each method (A) when the noise was added gradually (3-hour sampling frequency) and (B) when the sampling frequency was changed (noise level was 0.4), and both simulation datasets were not decaying. The *P*-value calculated by Tukey-Kramer test. The error bars indicate the standard deviations (n=3). MCC values and noise have no unit.

2.3.6 Comparison of MICOP and existing methods for noisy one-cycle data

I compared the accuracy of MICOP and existing methods for one-cycle data (Fig. 2.9). Among all conditions (method, noise, and sampling frequency), determination accuracies using one-cycle were lower than those using two cycles. All methods did not work under all conditions at the 4 hours sampling frequency. Meanwhile, MICOP and JTK showed high performances under sampling conditions $\leqq$ 3 hours.

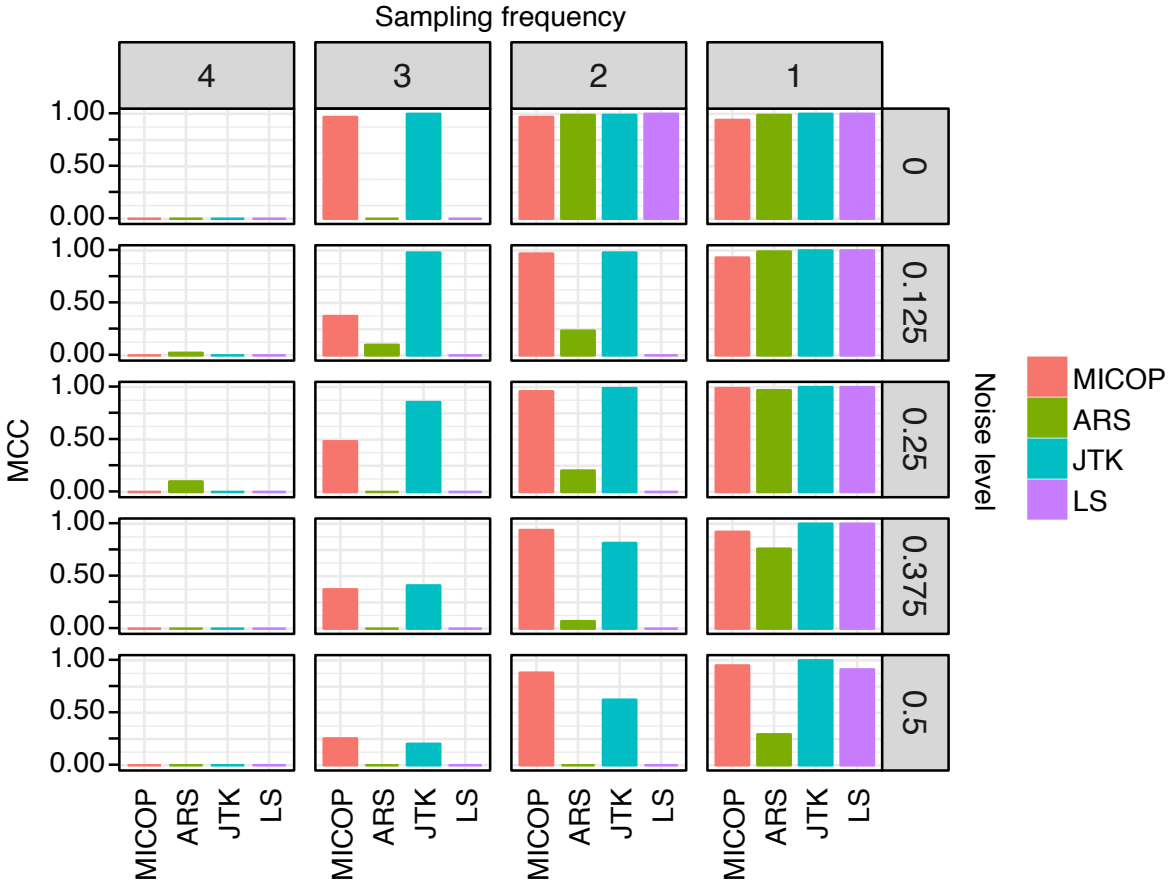

Fig. 2.9 Application of MICOP for one-cycle time-series data
Row order indicates noise level and column order indicates a sampling interval (hours). The colors refer to each method. Sampling intervals (hours) and noise level were gradually adjusted. Noise levels and MCC values have no units.

2.3.7 Phase estimation of MICOP and existing methods

I compared the phase estimation accuracy of MICOP and existing methods using time-shifted data containing noise and having a low sampling frequency without attenuation (Fig. 2.10). As a result, an absolute time lag of MICOP showed the smallest gaps. ARS, JTK, and LS showed average 0.75–1 hour, relatively MICOP showed 0.5 hours.

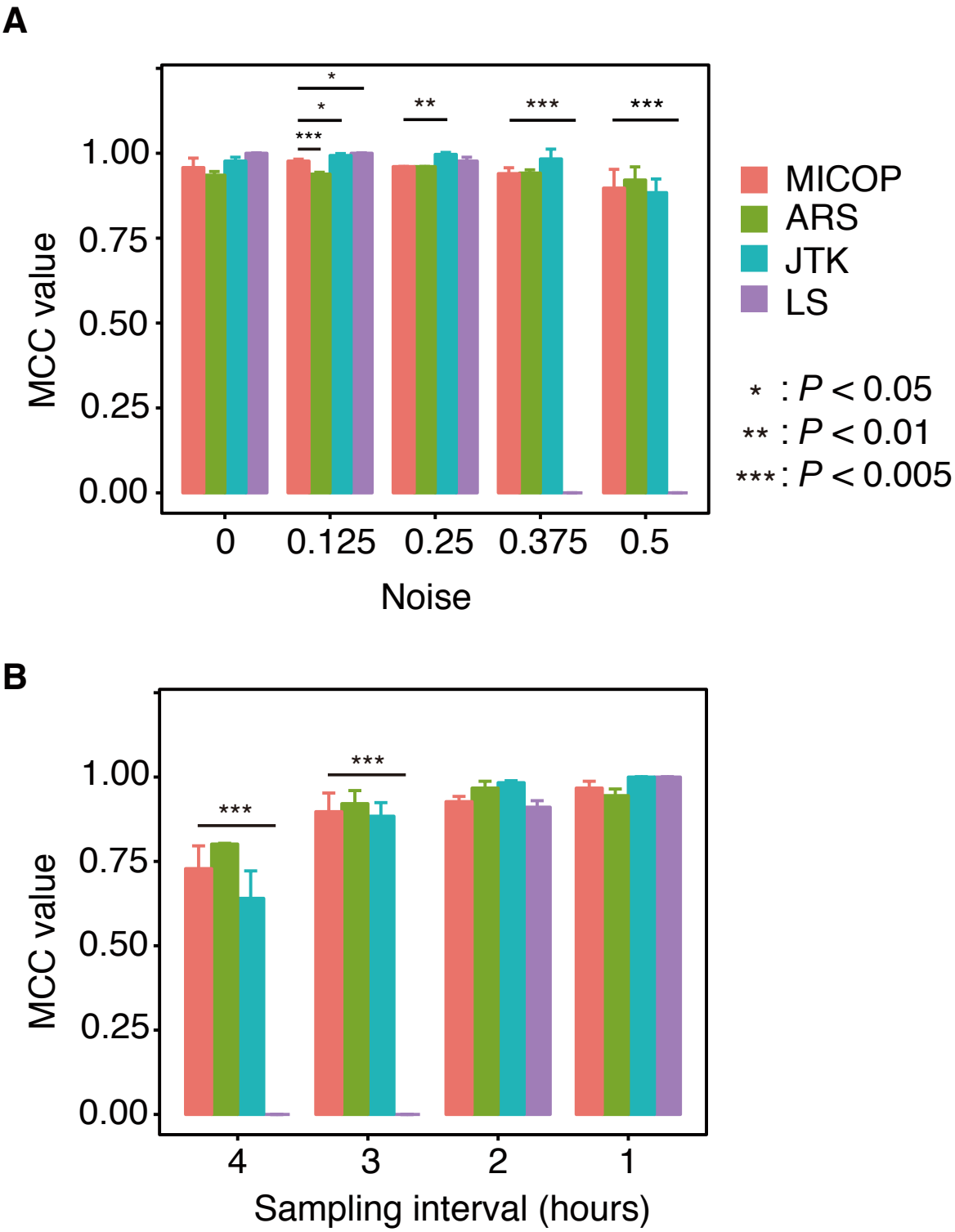

Fig. 2.10 Phase estimation of MICOP and existing methods for time-shifted data
(A) Histogram of time lags between correct phase and estimated phase. The x-axis indicates the frequency and the y-axis is a time lag (hours). (B) Absolute time lag (hours) between correct phase and estimation phase of MICOP and existing methods. *P*-values were calculated by Welch's *t*-test. The error bars indicate the standard deviations (n=1000).

2.3.8 Amplitude estimation of MICOP and existing methods

I compared the phase estimation accuracy of MICOP and existing methods using 1000 time-series data having a different amplitude (Fig. 2.11). The r^2 values of each method were more than 0.9, and there are no significant changes in each method.

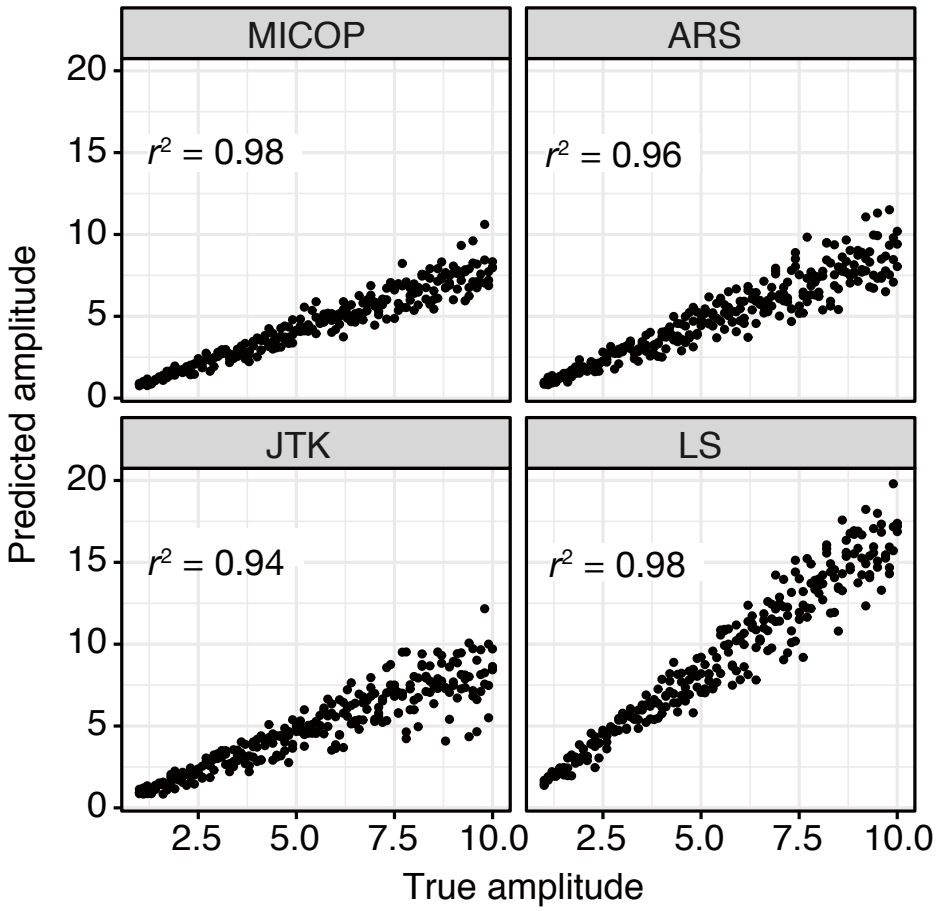

Fig. 2.11 Amplitude estimation of MICOP and existing methods for time-series data having different amplitude

The y-axis indicates true amplitude and x-axis indicates predicted amplitude. The r^2 values (Pearson correlation) were shown in each box. The true amplitude and predicted have no units.

2.3.9 Reanalysis of previously reported time-resolved proteomics datasets

I reanalyzed the time-series proteome data for mouse liver reported by Mauvoisin *et al*. using C57BL/6 and those reported by Robles *et al*. using C57BL/6J, as well as simulated data released by Wu *et al*. (Fig. 2.12, Tables 2.1, and 2.2) [40, 41, 54]. The numbers of significantly oscillating proteins assessed by standard harmonic regression were 9 (the F test for multilinear regression, $P < 0.01$), 9 (Fisher's exact test, $P < 0.01$), and 3 ($P < 0.01$) for biological data in the original work. Meanwhile, 32, 22, and 5 proteins were judged as being significantly oscillating for C57BL/6J, C57BL/6, and Wu's simulated data by MICOP, respectively ($P < 0.05$). The numbers of proteins judged to be significantly oscillating in both the original work and MICOP were 2, 8, and 2 for biological data, respectively. The numbers of proteins judged as being significantly oscillating for the three tests mentioned above only by MICOP were 30, 14, and 3 for biological data, respectively.

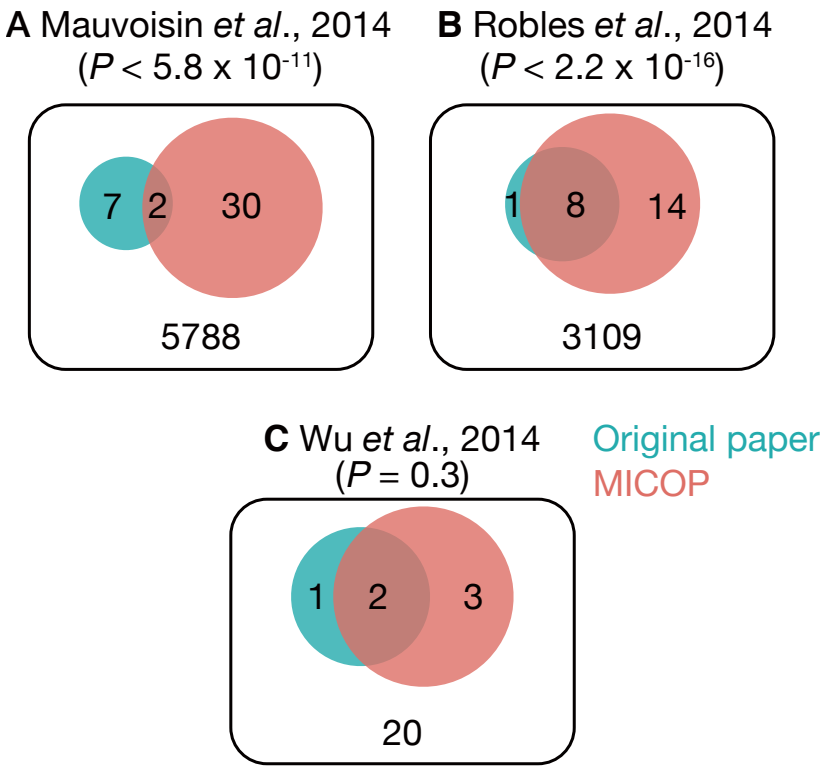

Fig. 2.12 Venn diagrams of significant molecules the levels of which oscillate
MICOP reanalyzed published time-resolved datasets, and Venn diagrams were constructed to quantify the overlap between MICOP and the original article. A and B represent mouse proteomics data: (A) C57BL/6J [41], (B) C57BL/6 [40], and (C) Wu's simulated data [54]. Blue indicates the original article and red indicates MICOP.

Table 2.1 Novel oscillating protein candidates detected by MICOP

Gene name	Species	Condition	Tissue	Reference
Acot8	-	-	-	-
Acox1	*Homo sapiens*	LL	blood	[62]
	Mus musculus	LD	liver/SCN	[63, 64]
Acsl5	*Mus musculus*	DD/LD	liver/SCN	[65]
	Mus musculus	LD	SCN	[63]
Akr1c14	*Mus musculus*	LD	liver	[66]
	Mus musculus	DD	cartilage tissue	[67]
Cbs	*Mus musculus*	DD/LD	liver	[68–71]
	Mus musculus	LD	SCN	[63]
	Homo sapiens	LL	blood	[62]
Cct8	*Mus musculus*	LD	SCN	[63]
Ces1b	-	-	-	-
Chid1	*Mus musculus*	DD/LD	liver/SCN	[72]
Cxadr	*Mus musculus*	DD/LD	liver/SCN	[68]
Cyp4f14	-	-	-	-
Gns	*Homo sapiens*	LL	blood	[62]
	Mus musculus	DD/LD	liver	[66]
Golgb1	*Mus musculus*	DD/LD	liver/SCN	[63, 66, 73]
Gpx3	-	-	-	-
Hars	*Mus musculus*	LD	liver	[66]
Hrg	-	-	-	-
Mfap4	-	-	-	-
Mug1	*Mus musculus*	DD	liver	[68]
Pdcd6	*Mus musculus*	LD	liver/SCN	[63]
Ptms	*Mus musculus*	LD	SCN	[63]
Safb	-	-	-	-
Serpina6	*Mus musculus*	DD/LD	liver	[68]
Sf3b2	*Mus musculus*	DD/LD	telogen epidermis	[72]
Slc9a3r1	*Mus musculus*	DD/LD	liver	[72]
Snrpd3	*Mus musculus*	LD	liver	[73]
Stk38	*Mus musculus*	DD	liver	[72]
	Mus musculus	LD	SCN	[63]
Tpr	*Mus musculus*	DD/LD	liver	[66, 72]
Txndc15	*Mus musculus*	DD/LD	liver	[72, 73]
	Mus musculus	LD	SCN	[63]
Ubl4a	*Mus musculus*	LD	SCN	[63]
	Mus musculus	DD	liver	[72]
Uox	*Mus musculus*	DD/LD	liver	[62, 65, 68, 73]
Ythdf2	*Mus musculus*	LD	liver	[73]

Light-dark (LD) stands for the daily 24-hour light-dark cycle, light-light (LL) stands for constant lightness, and dark-dark (DD) stands for constant darkness conditions. Hyphens indicate that I could not find previous consistent works which prove the mRNA oscillation.

Table 2.2 Novel oscillating protein candidates detected by MICOP

Gene name	Species	Condition	Tissue	Reference
Anp32e	*Mus musculus*	DD/LD	liver	[65, 66]
Anpep	-	-	-	-
Cgn	*Mus musculus*	LD	liver	[73]
Csde1	*Mus musculus*	LD	liver	[65]
	Mus musculus	DD	SCN	[63]
Enpp4	*Mus musculus*	LD	liver/anagen epidermis	[66, 72, 73]
Gnl2	*Mus musculus*	DD	hippocampus/liver	[65, 74]
	Mus musculus	LD	SCN	[63]
Ldhb	*Homo sapiens*	LL	blood	[62]
	Mus musculus	LD	anagen epidermis, SCN	[63, 72]
Numa1	*Mus musculus*	LD	liver	[66]
	Mus musculus	DD	cartilage tissue	[67]
Prdx2	*Mus musculus*	LD	SCN	[63]
Rnf114	-	-	-	-
Slc4a1	-	-	-	-
Slco1b2	*Mus musculus*	DD/LD	liver	[65, 66, 68]
Tomm70a	-	-	-	-
Vps26a	-	-	-	-

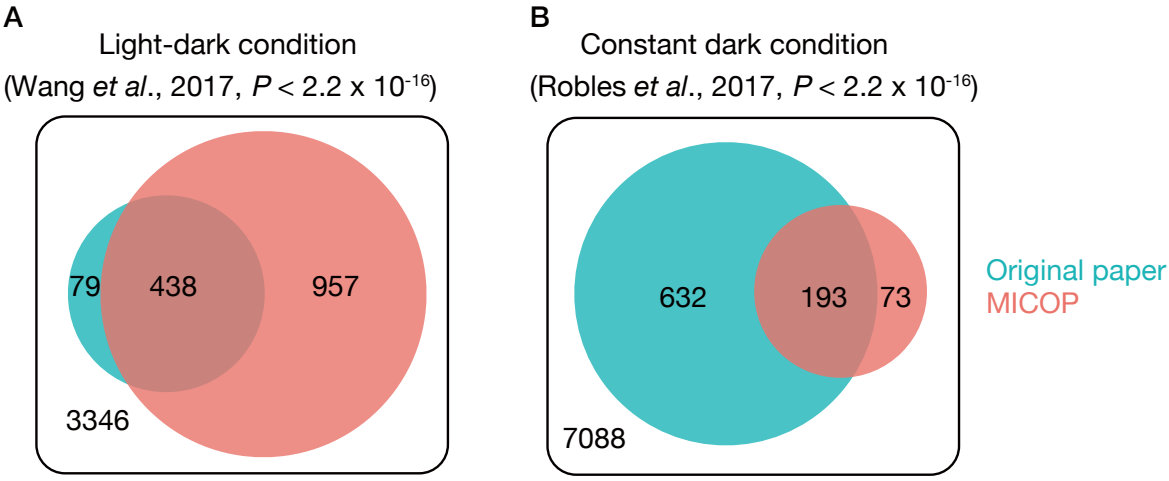

Fig. 2.13 Venn diagrams of significant molecules the levels of which oscillate of newly published datasets

MICOP reanalyzed newly published time-resolved datasets, and Venn diagrams were constructed to quantify the overlap between MICOP and the original article. (A) Light-dark condition with C57BL/6J [57], (B) dark-dark condition with C57BL/6 [61]. Blue indicates the original article and red indicates MICOP.

2.4 Discussion

Although many algorithms have been developed to extract molecules with rhythmic oscillation in their levels from large-scale time-series data derived from mass spectrometry systems or DNA sequencers, it is known that the accuracy and sensitivity of such methods depend on noise, sampling frequency, and waveform. In particular, the discussion of the prediction power in conditions of decaying oscillation is beneficial for not only bioinformaticians but also molecular biologists. In this research, I provide MICOP, which is classified as a time-domain method, and demonstrate that the algorithm is particularly effective for detecting decaying oscillation.

I compared the detection power of MICOP and previously reported algorithms for decaying oscillation. I revealed that regarding the power for detection of decaying oscillation, MICOP outperformed other algorithms (Fig. 2.5). In particular, MICOP showed a clear advantage when the sampling frequency was low. The reason is the MIC can effectively detect non-linear associations like associations between decaying oscillation and the reference sine curve (Fig. 2.1). ARS showed high performance following MICOP because de-trending at preprocessing seemed to cancel out the decay of time-series data. JTK had the third best detection power, although high performance was expected because it was based on Kendall's tau, which is a measure of rank correlation, and it did not depend on amplitude. The JTK output overly conservative *P*-values, therefore, the low threshold should improve performance evaluated by MCC. These results indicate that MICOP has the excellent performance for decaying oscillation, and suggests that the MIC-based approach that can detect non-linear associations is useful to detect decaying oscillation.

Moreover, I compared the MCC values for all methods of data containing gradual Gaussian noise to test the noise resistance (Fig. 2.8). As a result, MICOP showed equal performance to JTK and ARS in the range of the standard deviations of 0.125–0.500. This result indicated that the performance of MICOP for noisy data is equal to that of the existing

methods. This result suggests that the robustness to noise of MICOP is the same as that of well-known ARS and JTK, while the high performance of LS was limited to conditions with a low noise level. This numerical experiment revealed that the noise resistance of MICOP is the same as that of other widely used methods.

Clarifying the relationship between accuracy and sampling frequency in analyzing *omics* data, for which increasing the number of sampling points seems difficult, is important for determining the experimental design. As expected, with increasing sampling frequencies, the MCC values tended to increase (Figs. 2.5 and 2.8). The fact that the ARS, JTK, and LS could characterize oscillation and non-oscillation in almost all cases when the sampling interval was 2 hours or less is similar to the findings in the original research studies of various methods and research comparing them [43]. This result suggested that a high sampling frequency improved accuracy; therefore, sampling frequency should be as high as experimental constraints allow.

I applied MICOP and existing methods for one-cycle of data (Fig. 2.9). As expected, accuracy decreased for all methods when one-cycle was used. However, MICOP and JTK showed high MCC values among methods under this condition. Also, MICOP seems to outperform JTK under limited conditions which are low sampling frequency and high noise for one-cycle data (Fig. 2.9). Human *omics* data often have lower sampling frequencies, high noise levels, and only one-cycle. Our results suggest that MICOP and JTK have considerable potential for analyzing human *omics* datasets.

I reanalyzed the time-series proteomics data of C57BL/6J and C57BL/6 to test the performance of MICOP and explore additional candidates of proteins with a rhythmic change in their expression profiles [40, 41]. These datasets include the mouse liver proteome data obtained by sampling every 3 hours for 2 days, for which the analysis of the peptides was performed with a mass spectrometer. Approximately, 3000 protein types were detected in each study. Proteins that were detected in both MICOP and the original studies numbered

2 and 8 for C57BL/6J and C57BL/6, respectively (Fig. 2.12). This actual application for proteomics data suggests that MICOP can obtain results in a manner approximately similar to the existing methods. Specifically, the MICOP results were consistent with those in the original articles regarding these commonly identified proteins. Furthermore, the proteins those were uniquely identified with MICOP were numbered 30 and 14 for C57BL/6J and C57BL/6, respectively (Tables 2.1 and 2.2). These results strongly suggest that MICOP is a powerful tool to detect proteins with rhythmic changes in their expression levels from time-resolved proteomics data.

Although mass spectrometry-based approaches have been used for proteome-level studies of circadian rhythms, completely measuring mouse proteomes remains difficult. A comprehensive transcriptome analysis with parallel sequencers has revealed that 15–20 % of mouse liver mRNA significantly oscillates [75]. However, in these proteome studies of C57BL/6 and C57BL/6J, significantly oscillating protein are rare (< 1 % of detected total proteins; FDR < 0.05), a result inconsistent with those of mouse proteome studies. Multiple factors can explain this pattern. Typical clock proteins known as principle oscillators such as CRY1, CRY2, PER2, REV-ERBα, and CLOCK have comparatively low expression levels and are not detected in these studies [38]. Also, non-Gaussian experimental noise which is specific to MS measurement hampers the application of a statistical test on proteins [76]. These problems may be improved by analyzing higher quality proteome datasets with modern technologies [57, 61]. Some core circadian proteins such as CRY1, CRY2, PER2, REV-ERVα, and CLOCK could not be detected in the proteome studies which I refer in this paper [40, 41]. However, those proteins could be detected in recently published proteome datasets [57, 61]. Thus, the development of proteome analysis technology may resolve discrepancies between results of the transcriptome analysis and proteome analysis, and clarify connections within the circadian rhythm transcription and translation network.

I present a new list of proteins that oscillate by MICOP (Tables 2.1 and 2.2). The accuracy of these estimates is difficult to ascertain. Interestingly, a large fraction of candidates was presumed to oscillate in a previous transcriptome analysis [75]. Two independent expression patterns of genes encoding these proteins, I estimated that the proteins studies which measured both transcriptome and proteome of human samples revealed that only 30 % of the mRNA-protein correlation had statistically significant [77, 78]. This fact suggested that even if mRNA abundance is oscillating, protein abundance may not always be oscillating. However, about 90 % of mRNA-protein correlation showed positive. Hence rhythmic mRNA expression suggests the possibility of protein oscillation [77]. An overlap between re-analyzed proteomics data by MICOP and transcriptome analysis showed a consistent result.

MICOP accuracy tends to be low for data that do not perfectly fit a sine curve. The periodicity that MICOP can detect is subject to the shape of the reference curve, so changing the reference curve is necessary to detect asymmetric waveforms including sawtooth-like shapes like RAIN [76]. Furthermore, adjusting the FDR is essential for accurate prediction, since MICOP repeats the hypothesis tests. Also, verification with additional data such as periodic peak wave or overlapping sine wave is necessary in order to evaluate the accuracy of MICOP more precisely. Judgments of phase and cycle are possible in principle, but I did not perform them; therefore, this should be considered in future studies. Mutual information increased when the sample size was small, and a correlation between the two variables was null, even when the variables were random [79]. I solved this issue in MICOP by determining the *P*-value with the Monte Carlo method. When the time points (sample size) are small, the criterion for calculating the *P*-value increases, and when the time points are large, the criterion for calculating the *P*-value decreases (Fig. 2.14). In this paper, I presented MICOP, which is the MIC-based algorithm, for predicting periodic patterns in large-scale time-resolved protein expression profiles. The performance test using artificially generated

simulation data revealed that the performance of MICOP for decaying data was superior to that of the existing widely used methods. Additionally, I indicated that MICOP is compatible with noisy data obtained with a low sampling frequency. Furthermore, the performance test using actual mouse proteomics data suggested that MICOP may be able to provide novel findings from proteomics data. Specifically, it can reveal novel findings from time-series data and may contribute to biologically significant results. This study suggests that MICOP is an ideal approach for detecting and characterizing oscillations in time-resolved *omics* datasets.

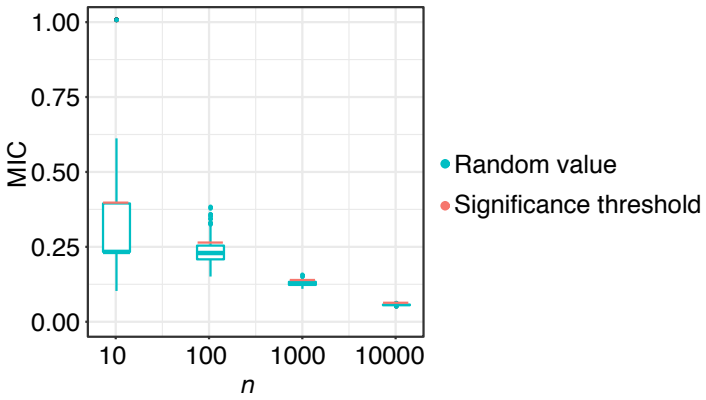

Fig. 2.14 Monte Carlo simulation to calculate *P*-values
MIC values were calculated between random numbers. The x-axis indicates sample number (n time points) and the y-axis indicates MIC. The box plots show the first quartile and third quartile range of the data and outliers. The blue color represents random values and the red color represents the significance threshold (5 %). MIC has no unit.

2.5 Conclusion

In this paper, I presented MICOP for predicting periodic patterns in large-scale time-resolved protein expression profiles. The performance test using artificially generated simulation data revealed that the performance of MICOP for decaying data was superior to that of the existing widely used methods. Additionally, I indicated that MICOP is compatible with noisy data obtained with a low sampling frequency. Furthermore, the performance test using actual mouse proteomics data suggested that MICOP may be able to provide novel findings from proteomics data. Specifically, it can reveal novel findings from time-series data and may contribute to biologically significant results. This study suggests that MICOP is an ideal approach for detecting and characterizing oscillations in time-resolved *omics* datasets.

Chapter 3

Amino acids acetylation by orphan enzyme YhhY in *Escherichia coli*

3.1 Background

Escherichia coli (*E. coli*) is one of the most studied model organisms. However, even in *E. coli*, the function of 30–40 % of all enzymes has not been experimentally demonstrated [80–86]. Therefore, there is also a large number of metabolites whose responsible enzymes have not been identified, which are referred to as "orphan metabolites" including some acetylated amino acids. These molecules create gaps in metabolic maps and represent an obstacle to understanding the metabolic systems [87].

Many algorithms have been developed to predict the functions of protein from their amino acid sequences, such as homology-based methods, structure-based methods, and genomic context-based methods [88–92]. Homology-based methods are based on the logic that similar sequences have similar functions and represent the most widely used prediction methods [88]. Structural similarity is a useful index for predicting protein function because 3D structures are sometimes better conserved than sequences are [90]. In prokaryotes, genes that are physically close on chromosomes have been conserved over the course of

evolution as operons and often tend to interact with each other [80, 91]. Therefore, the context-based method called the gene neighbor method using position information in the genome is particularly useful in prokaryotes such as *E. coli* [93]. However, these informatics-based approaches are only estimation methods, and experimental validation is required for definitive proof of a molecular function. The experimental validation is essential to obtain a deep understanding of metabolic systems.

Using mass spectrometry, I aimed to identify the activity of an uncharacterized enzyme of *E. coli*. In this context, it has been reported 40 years ago that *E. coli* extract acetylated phenylalanine, histidine, and alanine dependent on acetyl-CoA [94]. However, despite my investigations, I have not yet identified enzymes that acetylate these amino acids. I selected the uncharacterized enzyme YhhY in *E. coli*, as an example of enzyme characterization system using mass spectrometry. The function of YhhY was predicted to be acetyltransferase based on amino acid sequence the similarity, and the growth rate of *yhhY* knockout strain was suppressed in the presence of metal ions [95, 96]. However, the substrate of YhhY was not experimentally identified. Therefore, I aimed to establish enzyme characterization system with the functionally unknown enzymes YhhY of *E. coli* by combining a classical acetylation assay method and mass spectrometry-based metabolome analysis.

3.2 Materials and methods

3.2.1 Strains and culture conditions

The *E. coli* K-12 BW25113 strain was used as the wild-type, and the *yhhY* gene deletion mutants of the wild-type were obtained from the Keio collection [97, 98]. These strains were used for metabolome analysis. In addition, I used a *yhhY*-cloned strain obtained from the ASKA library for protein expression and metabolome analysis [99]. All the procedures for cultivating the wild-type and Δ*yhhY* mutant were performed using M9 medium (4.0 g

glucose, 6.8 g Na_2HPO_4, 3.0 g KH_2PO_4, 0.5 g NaCl, 1.0 g NH_4Cl per litter) supplemented with 2 mM $MgSO_4$, 100 μM $CaCl_2$, 0.01 % thiamine, and with or without 10 μM $CoCl_2$. The cultivation was carried out at 37 °C in a working volume of 50 mL in a 500 mL flask with shaking. The cultivation of the *yhhY*-overexpressing strain was performed with LB medium (10 g tryptone, 5 g yeast extract, 10 g NaCl per 1 liter). The spectrometer V-530 (JASCO, Tokyo, Japan) was used to measure OD600 nm for the normalization of cell number.

3.2.2 Expression and purification of recombinant proteins

Six × histidine-tagged YhhY was purified from ASKA clone [99]. Overnight Autoinduction System (Novagen, WI, USA) was used to perform the expression of the YhhY protein. Cells carrying the plasmid including the *yhhY* gene were inoculated in LB medium containing chloramphenicol for maintenance and incubated at 37 °C overnight for pre-culture. Two microliters of the saturated culture were inoculated into 2 ml of fresh LB medium supplemented with a solution for protein overexpression, following the manufacturer's instructions, and was cultured at 37 °C for 16 hours with shaking. Recombinant YhhY protein was purified using MagneHis Ni-Particles (Promega, WI, USA), following the manufacturer's instructions. Glycerol was added to a final concentration of 50 %, and the sample was stored at −20 °C until use.

3.2.3 The *in vitro* assay of YhhY

The reaction was performed in reaction buffer (50 mM Tricine, 20 mM KCl, 5 mM $MgCl_2$, 5 mM MnCl, 0.1 mM Ca^{2+}) supplemented with 20 mM acetyl-CoA, 10 mM 20 amino acids as substrates, 1 mM 2-morpholinoethanesulfonic acid, and 1 mM 3-aminopyrrolidine as internal standards for the MS analysis. The reaction buffer supplemented with 5 μL of beads with attached proteins was incubated at 37 °C overnight with shaking.

3.2.4 Extraction for metabolome analysis

Sample preparation was performed by a method established by Ohashi *et al*. [100]. An aliquot of 50 mL of culture broth (OD600 nm≈0.5) was filtered using a 0.45 μm pore size filter (GE Healthcare, IL, USA). The cells on the filter were washed twice with 20 mL of Milli-Q water at 37 °C, and the metabolism was stopped by steeping in 5 mL of methanol at 4 °C containing 2 μM 2-morpholinoethanesulfonic acid, 2 μM D-camphor-10-sulfonic acid, and 2 μM methionine sulfone as internal standards (Merck, Darmstadt, Germany). Sonication separated the cells on the filter. Four milliliters of methanol extract solution was stored at −80 °C until the next step. A total of 4 mL of chloroform and 1.6 mL of Milli-Q water were added to 4 mL of methanol extract solution, which was then fully mixed. The solution was centrifuged at 2,300 g for 5 min at 4 °C, and the 4 mL of the separated methanol layer was filtered by centrifugation at 2,300 g for 15 min at room temperature through a 5 kDa cutoff filter to remove high-molecular-weight compounds (Merck, Darmstadt, Germany). The filtrate was lyophilized and then dissolved in 50 μL Milli-Q water containing 1,3,5-benzenetricarbonyl and 3-aminopyrrolidine as internal standards for migration time normalization.

3.2.5 Instruments and conditions for CE-MS analysis

Metabolome analysis was performed using an Agilent CE Capillary Electrophoresis System equipped with an Agilent 6210 time-of-flight mass spectrometer, Agilent 1100 isocratic HPLC pump, Agilent G1603A capillary electrophoresis–mass spectrometry (CE-MS) adapter kit, and Agilent G1607A CE-electrospray ionization-MS sprayer kit (Agilent Technologies, CA, USA). The system was controlled by Agilent G2201AA ChemStation software for CE. CE-MS analysis was carried out using a modified version of the method described by Hirayama *et al*. [101], with two minor changes: sample solutions were injected at 50 mbar for 5 seconds, and the flow rate of nitrogen gas was set at 7 psig. All chemical standards

were dissolved in Milli-Q water, 0.1 M HCl or 0.1 M NaOH to obtain 1, 10, or 100 mmol/L stock solutions. The standard chemical mixture solution for the analysis was prepared by diluting with Milli-Q water to 20 μM each.

3.2.6 CE-MS data processing and statistical analysis

The in-house software for the identification and quantification of metabolites was used to process raw data. The data processing flow consisted of the following five steps: (1) noise filtering, (2) baseline correction, (3) migration time alignment, (4) peak detection, and (5) integration of a peak area. Obtained migration times of peaks in multiple samples were normalized using the dynamic programming-based techniques [102]. All the peak areas were divided by the area of internal standards for the normalization of sensitivity fluctuation among multiple measurements. Peaks under a threshold signal-to-noise ratio of 2 were considered undetected. The peaks derived from metabolites were identified if the measured m/z values and normalized migration times corresponded to standard compounds with an error tolerance of 40 ppm and 0.01 min. Processed peak lists were exported for further statistical analysis. The obtained metabolite concentration was normalized using OD600 nm.

3.3 Results

3.3.1 Amino acids acetylation of YhhY *in vitro*

My purpose is to reveal YhhY activity by traditional biochemical experiments and recently established techniques such as MS-based metabolome analysis. First, *E. coli* carrying a vector for *yhhY* gene overexpression was cultured, YhhY was subsequently purified, and a single band of the expected size was detected, but unsuspected bands were not detected (Fig. 3.1). Next, purified YhhY, the 20 amino acids, and acetyl-CoA were added as sources of acetyl groups into the test tube and reacted. Acetylation using acetyl-CoA produces free

CoA, and Ellman's reagent can be used to detect the byproducts by staining. Ellman's reagent which reacted with the thiol group has a maximum absorption wavelength at OD412 nm. In this experiment, OD440 nm was measured with a spectrometer to quantify free CoA. The results showed that free CoA was produced when YhhY was reacted with methionine, histidine, and phenylalanine (Fig. 3.2).

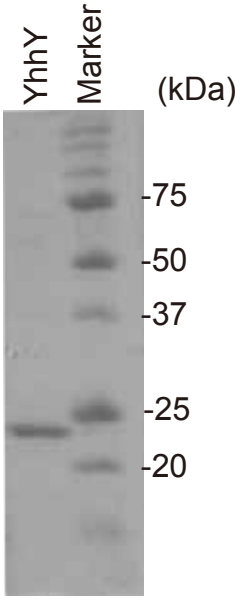

Fig. 3.1 SDS-PAGE of purified YhhY
Sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analyzed the purity of purified YhhY. The right column contains markers, and the left contains purified YhhY.

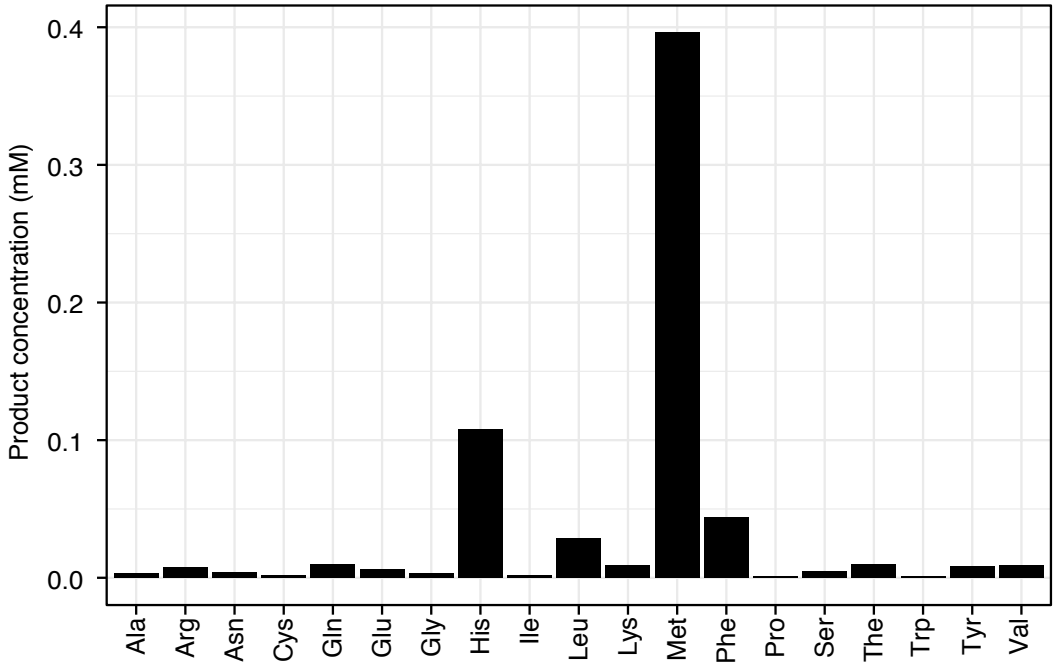

Fig. 3.2 Amino acids screening for YhhY target
The y-axis indicates product concentration, and the x-axis indicates each amino acid.

3.3.2 Inhibition by thiol reaction reagents, Lineweaver-Burk and ordination plots

Next, to clarify this acetylation event, an experiment involving inhibitor addition was performed. The addition of ethylmaleimide or iodoacetamide, well-known thiol reaction reagents competitively inhibiting acetylation in experimental systems, resulted in inhibition of the activity of YhhY (Fig. 3.3). Furthermore, a time-series experiment was performed in order to calculate the reaction rate, a Lineweaver-Burke plot and an ordination plot were drawn, and the rate of the reaction of YhhY with methionine was calculated (Fig. 3.4, Table 3.1). The Lineweaver-Burk plot revealed that acetylation by YhhY follows a ternary complex mechanism but not a ping-pong mechanism (Fig. 3.5).

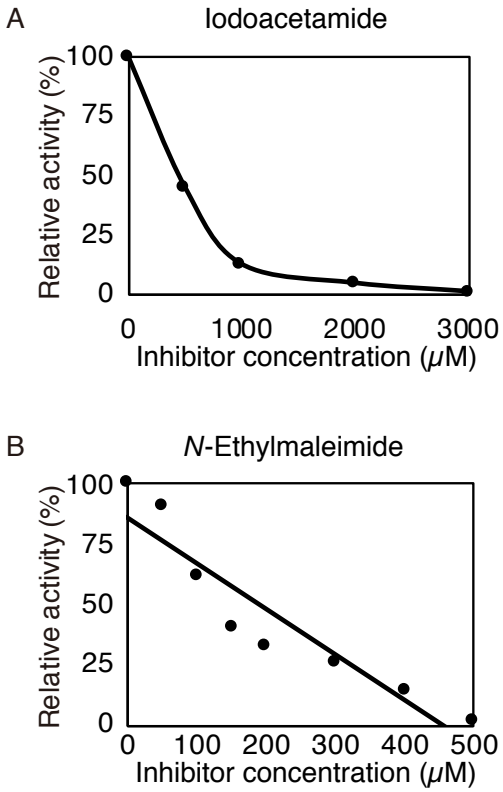

Fig. 3.3 Inhibition of acetylation event by *N*-ethylmaleimide and iodoacetamide
YhhY purified from *E. coli*, acetyl-CoA, and methionine were reacted with (A) iodoacetamide and (B) *N*-ethylmaleimide. The x-axis indicates inhibitor concentration (μM). The y-axis indicates relative activity (%) to that for the experiment without thiol reaction reagent.

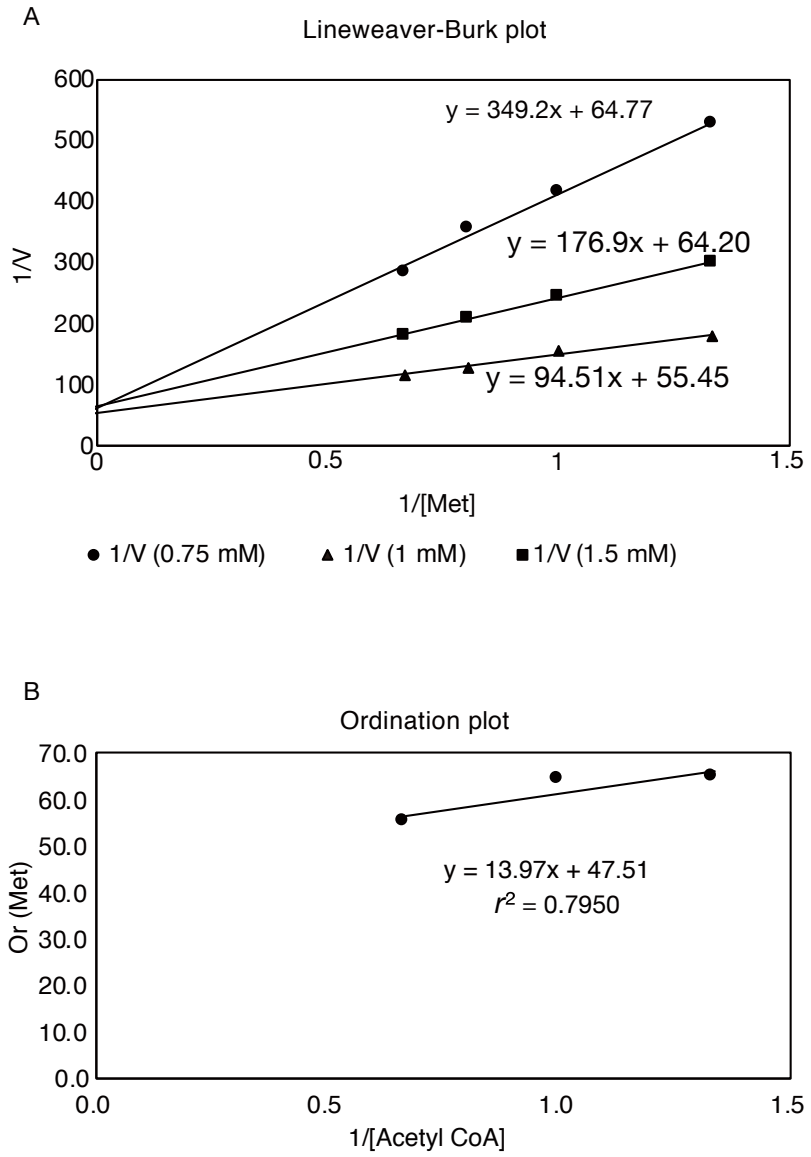

Fig. 3.4 Lineweaver-Burk plot and ordination plot for YhhY
(A) The Lineweaver–Burk plot and (B) the ordination plot for YhhY. The x-axis indicates the reciprocal of the methionine concentration and the y-axis indicates the reciprocal of the Lineweaver–Burk plot. The rhomboid represents 0.75 mM, acetyl-CoA, whereas rectangles represent 1 mM, and crosses represent 1.5 mM. The x-axis indicates the reciprocal of acetyl-CoA and the y-axis indicates the ordinary value of methionine.

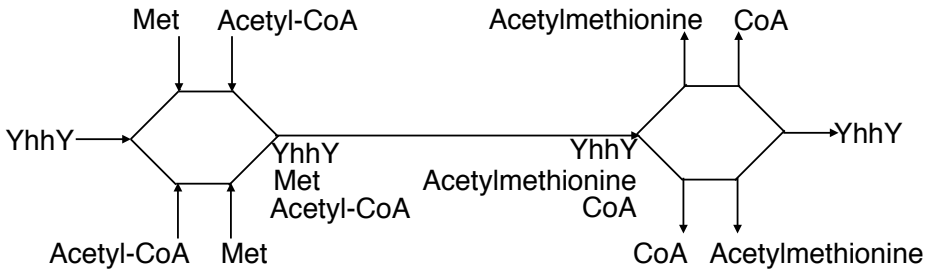

Fig. 3.5 Amino acids acetylation by YhhY
A schematic illustration of methionine acetylation catalyzed by YhhY. This acetylation event follows a ternary complex mechanism.

3.3.3 *In vitro* acetylation by extract of *yhhY*-overexpressing cells

Next, to prove that purified YhhY catalyzed these acetylation events, and not by a contaminating enzyme, I performed the *in vitro* experiment using extracts of *yhhY*-overexpressing cells. If a contaminated protein was responsible for the acetylation events, high-level acetylated products should also be detected in control extracts. The results showed that high-level acetylated products were detected only in the extract of *yhhY*-overexpressing cells (Fig. 3.6). Also, consistent with the purified YhhY experiment, the results showed different reaction rates for these three amino acids.

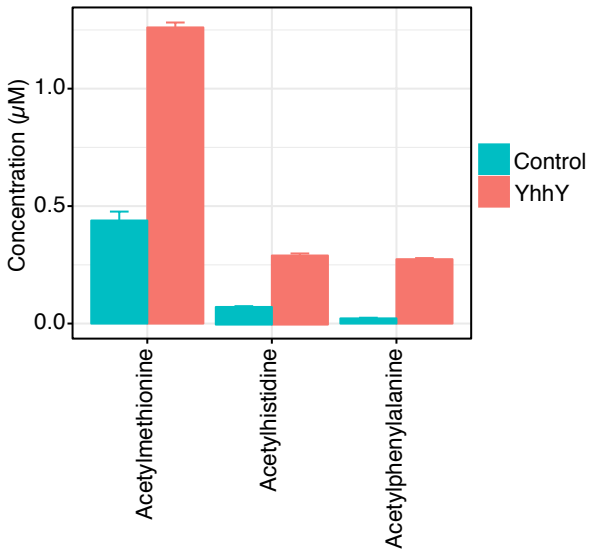

Fig. 3.6 Amino acids acetylation by extract of *yhhY*-overexpressing cells
Extracts of *yhhY*-overexpressing cells were incubated with acetyl-CoA and amino acids. Red bars indicate control and blue bars indicate samples containing extracts of *yhhY*-overexpressing cells. Error bars indicate standard deviations (n = 3).

3.3.4 Acetylated products in *yhhY*-overexpressing cells

Next, MS-based metabolome analysis of the *E. coli* strain carrying the *yhhY*-overexpressing vector was performed to clarify the intracellular activity of YhhY (Apx. A.1). First, CE-MS-based metabolome analysis revealed that acetylated amino acids were detected in *yhhY*-overexpressing cells only (Fig. 3.7). However, there were no significant changes in methionine, histidine, and phenylalanine.

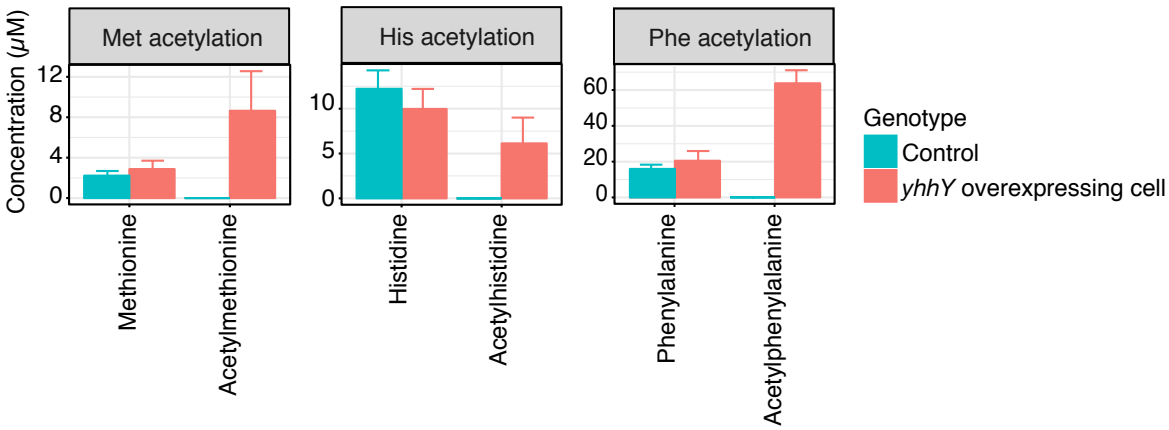

Fig. 3.7 Accumulation of acetylated amino acids in *yhhY*-overexpressing cells
Accumulation of methionine, histidine, phenylalanine, acetylmethionine, acetylhistidine, and acetylphenylalanine ($n = 3$) in control strain and *yhhY* gene-overexpressing strain. Error bars indicate standard deviation. The red bar is the control strain, the blue bar is the overexpressing strain.

3.3.5 Metabolome analysis of wild-type and *yhhY* deletion mutants under 10 μM cobalt conditions

Metabolome analysis of wild-type and *yhhY* deletion mutants under 10 μM cobalt conditions was performed to clarify *in vivo* activity of YhhY (Apx. A.2). The principal component analysis (PCA) showed metabolome level changes between the wild-type and *yhhY* deletion mutants (Fig. 3.8). The metabolome in wild-type cells is similar with and without cobalt in the medium; however, the metabolome is clearly different from the *yhhY* deletion mutants.

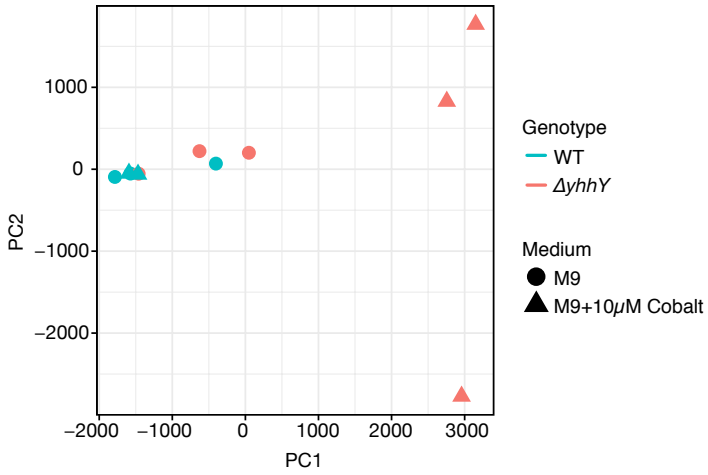

Fig. 3.8 PCA result of metabolome analysis with the wild-type and *yhhY* deletion mutants

Principal component (PC) analysis of *E. coli* metabolome. The color indicates genotype, and the shape indicates medium. Metabolome analysis was performed with CE-MS. WT indicates wild-type, and Δ*yhhY* indicates *yhhY* deletion mutants. The x-axis indicates PC1, and the y-axis indicates PC2. The contribution ratios were 72.1 and 21.3 % for PC1 and PC2, respectively.

3.4 Discussion

3.4.1 Specific amino acids acetylation by YhhY *in vitro*

The uncharacterized enzyme YhhY of *E. coli* was predicted to be as an amino acids acetyltransferase based on amino acid sequence homology; however, there was no experimental evidence of the function, and the target amino acids were unknown [87]. The *in vitro* screening of the 20 amino acids revealed that YhhY specifically acetylates methionine, histidine, and phenylalanine with different reaction rates (Fig. 3.2). Next, the well-known thiol-reactive reagents ethylmaleimide and iodoacetamide were shown to inhibit the YhhY activity in a concentration-dependent manner (Fig. 3.3). These results indicate that YhhY requires a free thiol group; thus, the acetylation of YhhY seems to involve a ternary complex mechanism and not a ping-pong mechanism.

3.4.2 Reaction rate and reaction pattern of YhhY

The time-course experiment to clarify the *in vitro* activity of YhhY was performed and the Lineweaver-Burk plot and an ordination plot were created to determine the substrate affinity of YhhY for acetyl-CoA and methionine, as shown in Fig. 3.4. Table 3.1 shows the K_{m} and V_{max} values. Based on the Lineweaver-Burk plot, when each substrate concentration is gradually changed, and each approximately straight line did not become parallel, it was suggested that the acetylation event by YhhY involves a ternary complex. In other words, YhhY binds to acetyl-CoA and methionine, and these three molecules from a complex; CoA and acetylated amino acids are then released as products. The fact that YhhY adopted a ternary covalent complex means that free CoA is necessary for the reaction, consistent with the inhibition of the reaction by thiol-reactive reagents such as ethylmaleimide and iodoacetamide. Fig. 3.5 shows a schematic diagram of this reaction model. I showed that YhhY is an acetyltransferase for methionine, histidine, and phenylalanine with different

reaction rates and obtained detailed kinetic parameters for methionine. Furthermore, I showed that this acetylation event follows a ternary complex mechanism.

Table 3.1 The reaction rate of YhhY with methionine

K_m for Met	K_m for acetyl-CoA	V_{max}
0.0047	0.29	0.021

3.4.3 The *in vitro* assay by *E. coli* extract

The *in vitro* assay using the extract of *yhhY*-overexpressing cells was performed to eliminate the possibility of a contaminating enzyme having acetylated the amino acids (Fig. 3.6). The acetylated amino acids were detected in *yhhY*-overexpressing cells, but not in the control sample. This result strongly supported that YhhY is responsible for the acetylation event.

3.4.4 *In vivo* activity of YhhY

Finally, a metabolome analysis was performed to clarify the *in vivo* function of YhhY in living cells. The results showed that acetylmethionine, acetylhistidine, and acetylphenylalanine were detected only from the *E. coli* carrying the *yhhY*-overexpressing vector, but not from the control group carrying the empty vector without the *yhhY* gene. This result was consistent with the results of the *in vitro* reaction and strongly support the possibility that YhhY is responsible for the acetylation events of three kinds of amino acids in the cell. There were no large differences in the concentrations of methionine, histidine, and phenylalanine as substrates in the two groups. This result may have been because the influence of the reaction involving enzymes other than YhhY was greater than the influence of YhhY on the accumulation of these amino acids. Additionally, *in vitro* experiments showed the highest abundance of acetylmethionine, but the highest abundance was acetylphenylalanine *in vivo*. I can assume two possible hypotheses to explain the differences between the results of *in vitro* and *in vivo*. One is the reaction rate inconsistency, and the other is intracellular acetylated products of living cells. In general, acetylation of amino acids and proteins has mainly been performed in eukaryotes, and acetylation is considered to be very limited in prokaryotes. However, in the recent research, more than 100 acetylated proteins were found in *E. coli* and Salmonella bacteria [103]. These results suggest that the acetylation events may play some roles also in prokaryotes. Many of the changes were observed in cells overexpressing YhhY suggesting that the reactions can occur under physiological conditions.

The concentrations of several other metabolites were significantly changed in overexpressing cells although their link to acetylation is not obvious. Some of the changes may have been indirect and may have been linked to changes in other enzymes triggered by the YhhY modification of some metabolites. YhhY has previously been shown to be activated by iron and thus to be part of the ferric uptake regulator (fur) response [104]. Also, it has been suggested to possibly play a role in revertants of an acetate toxic mutant [105]. It was reported that YhhY is overexpressed in biofilms (4.3 times) and that its mutation impaired growth on M63B1 medium (0.4 % glucose) containing thiamine and supplemented with iron [106]. YhhY appears to be activated by the fur response indirectly. Fur activates the expression of a small RNA that in turn is a repressor of YhhY. Fur activation thus results in increased YhhY expression. In a phenotypic screen of over 350 conditions on rich LB medium, cobalt treatment was shown to have a severe effect on growth and fitness [107]. For this reason, I chose to apply cobalt treatment as a way of highlighting differences between the wild-type and the *yhhY* knockout strain. I did not confirm the expression level of YhhY expression, but used 10 μM cobalt, a concentration that does not significantly affect growth compared with that in the wild-type, so that metabolomics can be performed without concern for growth rate-related effects on metabolite profiles.

3.4.5 Comparison of YhhY activity with previous reports

YhhY is probably the enzyme characterized in the study by Krishna *et al*. as it was reported to produce *N*-acetylphenylalanine and to have properties similar to YhhY [108]. The enzyme can acetylate both phenylalanine and histidine and has a molecular weight of 24 kDa. Also, the reported enzymes showed transferase activity for leucine and methionine.

These results were mostly consistent with those obtained by a recently published *in vitro* metabolomics approach for the enzyme discovery [109]. Sévin *et al*. incubated lysates of YhhY-overexpressing cells with substrates, suggesting that YhhY lysate acetylated

methionine and phenylalanine and proposed a new annotation for *yhhY* as an L-amino acid *N*-acetyltransferase (*aaaT*). Here, it should be noted that the acetylation of histidine that I detected was not previously reported in the study mentioned above. In my results, the acetylation event of histidine was weaker than the acetylation events of other amino acids. Also, in the previous work, the purification of YhhY in a soluble state was not achieved, nor was the function of YhhY estimated by reaction with cell extracts. In the cell extract assay, it is not possible to directly clarify the relationship between the overexpressed protein and the actual biochemical reaction. Also, I revealed the reaction mechanism and reaction rate of YhhY. These findings show that the sensitivity of mass spectrometry is essential for enzyme discovery by mass spectrometry-based metabolomics.

3.5 Conclusion

I aimed to establish an enzyme characterization system using mass spectrometry and selected the uncharacterized enzyme YhhY in *E. coli* as an example. YhhY was predicted to be an acetyltransferase; however, its substrates were not experimentally proven. *In vitro* screening using Ellman's reagent suggested that YhhY acetylated methionine, histidine, and phenylalanine. In addition, metabolome analysis using mass spectrometry of *in vitro* assay with the *E. coli* extract and *in vivo* cells supported the acetylation events by YhhY. Importantly, this result is consistent with another group's result [109]. Additionally, *in vivo* metabolome analysis of *yhhY* knockout suggested that YhhY canceled cobalt stress through a metabolic process. Thus, I established an enzyme characterization system using mass spectrometry, which can be used to characterize the function of other *E. coli* enzymes of unknown function.

Chapter 4

Concluding remarks

4.1 MICOP: Maximal information coefficient-based oscillation prediction

Circadian rhythms comprise oscillating molecular interactions, and the disruption of the homeostasis of which can cause various disorders. To understand this phenomenon, an accurate technique to identify oscillating molecules among *omics* datasets must be developed; however, this is still impeded by many difficulties, such as experimental noise and attenuated amplitude. To address these issues, I developed a new algorithm named maximal information coefficient-based oscillation prediction (MICOP), a sine curve-matching method. The performance of MICOP in labeling oscillation or non-oscillation was compared with those of three reported methods using MCC values. The numerical experiments were performed with time-series data with (1) mimicking of molecular oscillation decay and (2) high noise and low sampling frequency. The first experiment revealed that MICOP could accurately identify the rhythmicity of decaying molecular oscillation (MCC > 0.7). The second experiment revealed that MICOP was robust against high-level noise (MCC > 0.8) even upon the use of low sampling frequency data. As

an application, I utilized MICOP to analyze time-series proteome data of mouse liver. MICOP identified novel oscillating candidates numbered 14 and 30 for C57BL/6 and C57BL/6J, respectively. The performance test using artificially generated simulation data revealed that the performance of MICOP for decaying data was superior to that of widely used methods. It can reveal novel findings from time-series data and may contribute to biologically significant results. This study suggests that MICOP is an ideal approach for detecting and characterizing oscillations in time-resolved *omics* datasets.

4.2 Amino acids acetylation of orphan enzyme YhhY

Escherichia coli (*E. coli*) is one of the most studied model organisms. However, even in *E. coli*, for 30–40% of all enzymes, the functions have not been experimentally demonstrated. Therefore, there are also a large number of metabolites whose responsible enzymes have not been identified, which are known as "orphan metabolites" including some acetylated amino acids. These molecules create gaps in metabolic maps and have become an obstacle to understanding metabolic systems. To accelerate filling these gaps, I aimed to develop an enzyme characterization strategy based on mass spectrometry. I selected the uncharacterized enzyme YhhY of *E. coli* to evaluate the enzyme characterization system using a combination of the classical acetylation assay method and metabolome analysis. Screening using Ellman's reagent strongly suggested that YhhY specifically acetylates methionine, histidine, and phenylalanine. Next, mass spectrometry-based metabolome analysis suggested that YhhY, not a contaminated enzyme, was responsible for the acetylation event. Also, YhhY potentially has a relationship with metal-induced stress *in vivo*. The biochemical experiments revealed that the kinetic parameters and the acetylation by YhhY follow the ternary complex mechanism but not the ping-pong mechanism. In this thesis, I developed the enzyme characterization system using mass spectrometry, and the application of the system for uncharacterized

enzyme YhhY demonstrated the availability of the enzyme characterization strategy using mass spectrometry.

4.3 Mass spectrometry for systems biology

Large-scale experiments such as functional screening by mass spectrometry have produced huge data. The big issue of systems biology is how to extract biological significance in the large datasets. In this thesis, I aimed to develop new strategies to extract biological significance in *omics* data measured by especially mass spectrometry. First, I developed a novel algorithm to detect rhythmic molecules in *omics* data, MICOP: Maximal information coefficient-based oscillation prediction. Next, I established an enzyme characterization system using mass spectrometry and revealed the activity of uncharacterized enzyme YhhY in *E. coli*. These approaches using mass spectrometry data should contribute to a better understanding of the life systems.

References

[1] M. K. Chandrashekaran. Biological rhythms research: A personal account. *J Biosci*, 23(5):545–55, 1998.

[2] R. J. Konopka and S. Benzer. Clock mutants of drosophila melanogaster. *Proc Natl Acad Sci U S A*, 68(9):2112–6, 1971.

[3] P. Reddy, W. A. Zehring, D. A. Wheeler, V. Pirrotta, C. Hadfield, J. C. Hall, and M. Rosbash. Molecular analysis of the period locus in drosophila melanogaster and identification of a transcript involved in biological rhythms. *Cell*, 38(3):701–10, 1984.

[4] W. A. Zehring, D. A. Wheeler, P. Reddy, R. J. Konopka, C. P. Kyriacou, M. Rosbash, and J. C. Hall. P-element transformation with period locus DNA restores rhythmicity to mutant, arrhythmic drosophila melanogaster. *Cell*, 39(2):369–76, 1984.

[5] T. A. Bargiello, F. R. Jackson, and M. W. Young. Restoration of circadian behavioural rhythms by gene transfer in drosophila. *Nature*, 312(5996):752–4, 1984.

[6] F. K. Stephan and I. Zucker. Circadian rhythms in drinking behavior and locomotor activity of rats are eliminated by hypothalamic lesions. *Proc Natl Acad Sci U S A*,

69(6):1583–6, 1972.

[7] R. Y. Moore and V. B. Eichler. Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Res*, 42(1):201–6, 1972.

[8] M. H. Vitaterna, D. P. King, A. M. Chang, J. M. Kornhauser, P. L. Lowrey, J. D. McDonald, W. F. Dove, L. H. Pinto, F. W. Turek, and J. S. Takahashi. Mutagenesis and mapping of a mouse gene, clock, essential for circadian behavior. *Science*, 264(5159):719–25, 1994.

[9] D. P. King, Y. Zhao, A. M. Sangoram, L. D. Wilsbacher, M. Tanaka, M. P. Antoch, T. D. Steeves, M. H. Vitaterna, J. M. Kornhauser, P. L. Lowrey, F. W. Turek, and J. S. Takahashi. Positional cloning of the mouse circadian clock gene. *Cell*, 89(4):641–53, 1997.

[10] M. P. Antoch, E. J. Song, A. M. Chang, M. H. Vitaterna, Y. Zhao, L. D. Wilsbacher, A. M. Sangoram, D. P. King, L. H. Pinto, and J. S. Takahashi. Functional identification of the mouse circadian clock gene by transgenic bac rescue. *Cell*, 89(4):655–67, 1997.

[11] D. A. Paranjpe and V. K. Sharma. Evolution of temporal order in living organisms. *J Circadian Rhythms*, 3(1):7, 2005.

[12] Y. Shinohara, Y. M. Koyama, M. Ukai-Tadenuma, T. Hirokawa, M. Kikuchi, R. G. Yamada, H. Ukai, H. Fujishima, T. Umehara, K. Tainaka, and H. R. Ueda. Temperature-sensitive substrate and product binding underlie temperature-compensated Phosphorylation in the clock. *Mol Cell*, 67(5):783–798, 2017.

[13] J. S. Takahashi, H. K. Hong, C. H. Ko, and E. L. McDearmon. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nat Rev Genet*, 9(10):764–75, 2008.

[14] Y. Minami, T. Kasukawa, Y. Kakazu, M. Iigo, M. Sugimoto, S. Ikeda, A. Yasui, G. T. van der Horst, T. Soga, and H. R. Ueda. Measurement of internal body time by blood metabolomics. *Proc Natl Acad Sci U S A*, 106(24):9890–5, 2009.

[15] K. L. Eckel-Mahan, V. R. Patel, R. P. Mohney, K. S. Vignola, P. Baldi, and P. Sassone-Corsi. Coordination of the transcriptome and metabolome by the circadian clock. *Proc Natl Acad Sci U S A*, 109(14):5541–6, 2012.

[16] R. Dallmann, A. U. Viola, L. Tarokh, C. Cajochen, and S. A. Brown. The human circadian metabolome. *Proc Natl Acad Sci U S A*, 109(7):2625–9, 2012.

[17] E. D. Weitzman, D. Fukushima, C. Nogeire, H. Roffwarg, T. F. Gallagher, and L. Hellman. Twenty-four hour pattern of the episodic secretion of cortisol in normal subjects. *J Clin Endocrinol Metab*, 33(1):14–22, 1971.

[18] D. J. Kennaway, A. Voultsios, T. J. Varcoe, and R. W. Moyer. Melatonin in mice: rhythms, response to light, adrenergic stimulation, and metabolism. *Am J Physiol Regul Integr Comp Physiol*, 282(2):R358–65, 2002.

[19] F. Halberg. Chronobiology. *Annu Rev Physiol*, 31:675–725, 1969.

[20] A. Reinberg, M. Smolensky, and F. Levi. Aspects of clinical chronopharmacology. *Cephalalgia*, 3 Suppl 1:69–78, 1983.

[21] A. Reinberg and F. Halberg. Circadian chronopharmacology. *Annu Rev Pharmacol*, 11:455–92, 1971.

[22] V. Bocci. Administration of interferon at night may increase its therapeutic index. *Cancer Drug Deliv*, 2(4):313–8, 1985.

[23] B. Lemmer, B. Scheidel, and S. Behne. Chronopharmacokinetics and chronopharmacodynamics of cardiovascular active drugs. propranolol, organic nitrates, nifedipine. *Ann N Y Acad Sci*, 618:166–81, 1991.

[24] S. Ohdo, S. Koyanagi, H. Suyama, S. Higuchi, and H. Aramaki. Changing the dosing schedule minimizes the disruptive effects of interferon on clock function. *Nat Med*, 7(3):356–60, 2001.

[25] W. J. Hrushesky. Circadian timing of cancer chemotherapy. *Science*, 228(4695):73–5, 1985.

[26] S. Hasan, N. Santhi, A. S. Lazar, A. Slak, J. Lo, M. von Schantz, S. N. Archer, J. D. Johnston, and D. J. Dijk. Assessment of circadian rhythms in humans: comparison of real-time fibroblast reporter imaging with plasma melatonin. *FASEB J*, 26(6):2414–23, 2012.

[27] M. R. Smith and C. I. Eastman. Night shift performance is improved by a compromise circadian phase position: study 3. circadian phase after 7 night shifts with an intervening weekend off. *Sleep*, 31(12):1639–45, 2008.

[28] K. P. Jr. Wright, C. Gronfier, J. F. Duffy, and C. A. Czeisler. Intrinsic period and light intensity determine the phase relationship between melatonin and sleep in humans. *J Biol Rhythms*, 20(2):168–77, 2005.

[29] T. S. Horowitz, B. E. Cade, J. M. Wolfe, and C. A. Czeisler. Efficacy of bright light and sleep/darkness scheduling in alleviating circadian maladaptation to night work. *Am J Physiol Endocrinol Metab*, 281(2):E384–91, 2001.

[30] G. Labrecque and P. M. Belanger. Biological rhythms in the absorption, distribution, metabolism and excretion of drugs. *Pharmacol Ther*, 52(1):95–107, 1991.

[31] T. Kasukawa, M. Sugimoto, A. Hida, Y. Minami, M. Mori, S. Honma, K. Honma, K. Mishima, T. Soga, and H. R. Ueda. Human blood metabolite timetable indicates internal body time. *Proc Natl Acad Sci U S A*, 109(37):15036–41, 2012.

[32] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316(5824):593–7, 2007.

[33] S. Timm, A. Florian, M. Wittmiss, K. Jahnke, M. Hagemann, A. R. Fernie, and H. Bauwe. Serine acts as a metabolic signal for the transcriptional control of photorespiration-related genes in arabidopsis. *Plant Physiol*, 162(1):379–89, 2013.

[34] M. Sugimoto, D. T. Wong, A. Hirayama, T. Soga, and M. Tomita. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics*, 6(1):78–95, 2010.

[35] A. Diez-Noguera. Methods for serial analysis of long time series in the study of biological rhythms. *J Circadian Rhythms*, 11(1):7, 2013.

[36] M. E. Hughes, K. C. Abruzzi, R. Allada, R. Anafi, A. B. Arpat, G. Asher, P. Baldi, C. de Bekker, D. Bell-Pedersen, J. Blau, S. Brown, M. F. Ceriani, Z. Chen, J. C. Chiu, J. Cox, A. M. Crowell, J. P. DeBruyne, D. J. Dijk, L. DiTacchio, F. J. Doyle, G. E. Duffield, J. C. Dunlap, K. Eckel-Mahan, K. A. Esser, G. A. FitzGerald, D. B. Forger, L. J. Francey, Y. H. Fu, F. Gachon, D. Gatfield, P. de Goede, S. S. Golden, C. Green, J. Harer, S. Harmer, J. Haspel, M. H. Hastings, H. Herzel, E. D. Herzog, C. Hoffmann, C. Hong, J. J. Hughey, J. M. Hurley, H. O. de la Iglesia, C. Johnson, S. A. Kay, N. Koike, K. Kornacker, A. Kramer, K. Lamia, T. Leise, S. A. Lewis, J. Li, X. Li, A. C. Liu, J. J. Loros, T. A. Martino, J. S. Menet, M. Merrow, A. J. Millar, T. Mockler, F. Naef, E. Nagoshi, M. N. Nitabach, M. Olmedo, D. A. Nusinow, L. J. Ptacek, D. Rand, A. B. Reddy, M. S. Robles, T. Roenneberg, M. Rosbash, M. D. Ruben, S. S. C. Rund, A. Sancar, P. Sassone-Corsi, A. Sehgal, S. Sherrill-Mix, D. J. Skene, K. F. Storch, J. S. Takahashi, H. R. Ueda, H. Wang, C. Weitz, P. O. Westermark, H. Wijnen, Y. Xu, G. Wu, S. H. Yoo, M. Young, E. E. Zhang, T. Zielinski, and J. B. Hogenesch. Guidelines for genome-scale analysis of biological rhythms. *J Biol Rhythms*, 32(5):380–93, 2017.

[37] J. A. Mohawk, C. B. Green, and J. S. Takahashi. Central and peripheral circadian clocks in mammals. *Annu Rev Neurosci*, 35:445–62, 2012.

[38] N. Koike, S. H. Yoo, H. C. Huang, V. Kumar, C. Lee, T. K. Kim, and J. S. Takahashi. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105):349–54, 2012.

[39] C. L. Partch, C. B. Green, and J. S. Takahashi. Molecular architecture of the mammalian circadian clock. *Trends Cell Biol*, 24(2):90–9, 2014.

[40] M. S. Robles, J. Cox, and M. Mann. *In-vivo* quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet*, 10(1):e1004047, 2014.

[41] D. Mauvoisin, J. Wang, C. Jouffe, E. Martin, F. Atger, P. Waridel, M. Quadroni, F. Gachon, and F. Naef. Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proc Natl Acad Sci U S A*, 111(1):167–72, 2014.

[42] D. Ono, K. Honma, and S. Honma. Circadian and ultradian rhythms of clock gene expression in the suprachiasmatic nucleus of freely moving mice. *Sci Rep*, 5:12310, 2015.

[43] A. Deckard, R. C. Anafi, J. B. Hogenesch, S. B. Haase, and J. Harer. Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics*, 29(24):3174–80, 2013.

[44] M. Ukai-Tadenuma, R. G. Yamada, H. Xu, J. A. Ripperger, A. C. Liu, and H. R. Ueda. Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell*, 144(2):268–81, 2011.

[45] D. Chudova, A. Ihler, K. K. Lin, B. Andersen, and P. Smyth. Bayesian detection of non-sinusoidal periodic patterns in circadian expression data. *Bioinformatics*, 25(23):3114–20, 2009.

[46] M. Straume. DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods Enzymol*, 383:149–66, 2004.

[47] M. E. Hughes, J. B. Hogenesch, and K. Kornacker. JTK_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms*, 25(5):372–80, 2010.

[48] S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 2004.

[49] R. Takalo, H. Hytti, and H. Ihalainen. Tutorial on univariate autoregressive spectral analysis. *J Clin Monit Comput*, 19(6):401–10, 2005.

[50] R. Yang and Z. Su. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12):i168–74, 2010.

[51] J. D. Levine, P. Funes, H. B. Dowse, and J. C. Hall. Signal analysis of behavioral and molecular cycles. *BMC Neurosci*, 3:1, 2002.

[52] C. J. Langmead, A. K. Yan, C. R. McClung, and B. R. Donald. Phase-independent rhythmic analysis of genome-wide expression patterns. *J Comput Biol*, 10(3-4):521–36, 2003.

[53] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–24, 2011.

[54] G. Wu, J. Zhu, J. Yu, L. Zhou, J. Z. Huang, and Z. Zhang. Evaluation of five methods for genome-wide circadian gene identification. *J Biol Rhythms*, 29(4):231–42, 2014.

[55] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[56] C. Vollmers, R. J. Schmitz, J. Nathanson, G. Yeo, J. R. Ecker, and S. Panda. Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell Metab*, 16(6):833–45, 2012.

[57] J. Wang, D. Mauvoisin, E. Martin, F. Atger, A. N. Galindo, L. Dayon, F. Sizzano, A. Palini, M. Kussmann, P. Waridel, M. Quadroni, V. Dulic, F. Naef, and F. Gachon. Nuclear proteomics uncovers diurnal regulatory landscapes in mouse liver. *Cell Metab*, 25(1):102–117, 2017.

[58] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[59] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–62, 1976.

[60] H. P. Van Dongen, E. Olofsen, J. H. VanHartevelt, and E. W. Kruyt. Searching for biological rhythms: peak detection in the periodogram of unequally spaced data. *J Biol Rhythms*, 14(6):617–20, 1999.

[61] M. S. Robles, S. J. Humphrey, and M. Mann. Phosphorylation is a central mechanism for circadian control of metabolism and physiology. *Cell Metab*, 25(1):118–27, 2017.

[62] C. S. Moller-Levet, S. N. Archer, G. Bucca, E. E. Laing, A. Slak, R. Kabiljo, J. C. Lo, N. Santhi, M. von Schantz, C. P. Smith, and D. J. Dijk. Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proc Natl Acad Sci U S A*, 110(12):E1132–41, 2013.

[63] W. G. Pembroke, A. Babbs, K. E. Davies, C. P. Ponting, and P. L. Oliver. Temporal transcriptomics suggest that twin-peaking genes reset the clock. *Elife*, 4:e10518, 2015.

[64] B. C. Goh, X. Wu, A. E. Evans, M. L. Johnson, M. R. Hill, and J. M. Gimble. Food entrainment of circadian gene expression altered in PPAR$\alpha^{-/-}$ brown fat and heart. *Biochem Biophys Res Commun*, 360(4):828–33, 2007.

[65] M. E. Hughes, L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda, and J. B. Hogenesch. Harmonics of circadian gene transcription in mammals. *PLoS Genet*, 5(4):e1000442, 2009.

[66] J. L. Barclay, J. Husse, B. Bode, N. Naujokat, J. Meyer-Kovac, S. M. Schmid, H. Lehnert, and H. Oster. Circadian desynchrony promotes metabolic disruption in a mouse model of shiftwork. *PLoS One*, 7(5):e37150, 2012.

[67] N. Gossan, L. Zeef, J. Hensman, A. Hughes, J. F. Bateman, L. Rowley, C. B. Little, H. D. Piggins, M. Rattray, R. P. Boot-Handford, and Q. J. Meng. The circadian clock in murine chondrocytes regulates genes controlling key aspects of cartilage homeostasis. *Arthritis Rheum*, 65(9):2334–45, 2013.

[68] B. H. Miller, E. L. McDearmon, S. Panda, K. R. Hayes, J. Zhang, J. L. Andrews, M. P. Antoch, J. R. Walker, K. A. Esser, J. B. Hogenesch, and J. S. Takahashi. Circadian and clock-controlled regulation of the mouse transcriptome and cell proliferation. *Proc Natl Acad Sci U S A*, 104(9):3342–7, 2007.

[69] H. Tsuchiya, K. A. da Costa, S. Lee, B. Renga, H. Jaeschke, Z. Yang, S. J. Orena, M. J. Goedken, Y. Zhang, B. Kong, M. Lebofsky, S. Rudraiah, R. Smalling, G. Guo, S. Fiorucci, S. H. Zeisel, and L. Wang. Interactions between nuclear receptor shp and foxa1 maintain oscillatory homocysteine homeostasis in mice. *Gastroenterology*, 148(5):1012–23, 2015.

[70] P. Janich, A. B. Arpat, V. Castelo-Szekely, M. Lopes, and D. Gatfield. Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. *Genome Res*, 25(12):1848–59, 2015.

[71] C. Vollmers, R. J. Schmitz, J. Nathanson, G. Yeo, J. R. Ecker, and S. Panda. Circadian oscillations of protein-coding and regulatory rnas in a highly dynamic mammalian liver epigenome. *Cell Metab*, 16(6):833–45, 2012.

[72] M. Geyfman, V. Kumar, Q. Liu, R. Ruiz, W. Gordon, F. Espitia, E. Cam, S. E. Millar, P. Smyth, A. Ihler, J. S. Takahashi, and B. Andersen. Brain and muscle arnt-like protein-1 (BMAL1) controls circadian cell proliferation and susceptibility to uvb-induced DNA damage in the epidermis. *Proc Natl Acad Sci U S A*, 109(29):11758–63, 2012.

[73] K. L. Eckel-Mahan, V. R. Patel, S. de Mateo, R. Orozco-Solis, N. J. Ceglia, S. Sahar, S. A. Dilag-Penilla, K. A. Dyar, P. Baldi, and P. Sassone-Corsi. Reprogramming of the circadian clock by nutritional challenge. *Cell*, 155(7):1464–78, 2013.

[74] J. Renaud, F. Dumont, M. Khelfaoui, S. R. Foisset, F. Letourneur, T. Bienvenu, O. Khwaja, O. Dorseuil, and P. Billuart. Identification of intellectual disability genes showing circadian clock-dependent expression in the mouse hippocampus. *Neuroscience*, 308:11–50, 2015.

[75] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci U S A*, 111(45):16219–24, 2014.

[76] P. F. Thaben and P. O. Westermark. Detecting rhythms in time series with rain. *J Biol Rhythms*, 29(6):391–400, 2014.

[77] M. Gry, R. Rimini, S. Stromberg, A. Asplund, F. Ponten, M. Uhlen, and P. Nilsson. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, 10:365, 2009.

[78] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, and Cptac Nci. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–7, 2014.

[79] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl 2:S231–40, 2002.

[80] T. Yamada, A. S. Waller, J. Raes, A. Zelezniak, N. Perchat, A. Perret, M. Salanoubat, K. R. Patil, J. Weissenbach, and P. Bork. Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Mol Syst Biol*, 8:581, 2012.

[81] L. Chen and D. Vitkup. Distribution of orphan metabolic activities. *Trends Biotechnol*, 25(8):343–8, 2007.

[82] Y. Pouliot and P. D. Karp. A survey of orphan enzyme activities. *BMC Bioinformatics*, 8:244, 2007.

[83] A. D. Hanson, A. Pribat, J. C. Waller, and V. de Crecy-Lagard. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem J*, 425(1):1–11, 2009.

[84] M. Sorokina, M. Stam, C. Medigue, O. Lespinet, and D. Vallenet. Profiling the orphan enzymes. *Biol Direct*, 9:10, 2014.

[85] M. Y. Galperin and E. V. Koonin. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res*, 32(18):5452–63, 2004.

[86] L. Jaroszewski, Z. Li, S. S. Krishna, C. Bakolitsa, J. Wooley, A. M. Deacon, I. A. Wilson, and A. Godzik. Exploration of uncharted regions of the protein universe. *PLoS Biol*, 7(9):e1000205, 2009.

[87] M. Arita. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci U S A*, 101(6):1543–7, 2004.

[88] G. R. Reeck, C. de Haen, D. C. Teller, R. F. Doolittle, W. M. Fitch, R. E. Dickerson, P. Chambon, A. D. McLachlan, E. Margoliash, T. H. Jukes, and et al. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, 50(5):667, 1987.

[89] R. D. Sleator and P. Walsh. An overview of in silico protein function prediction. *Arch Microbiol*, 192(3):151–5, 2010.

[90] J. C. Whisstock and A. M. Lesk. Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, 36(3):307–40, 2003.

[91] T. Gabaldon and M. A. Huynen. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 61(7-8):930–44, 2004.

[92] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–6, 2000.

[93] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–901, 1999.

[94] A. Platt, H. C. Ross, S. Hankin, and R. J. Reece. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. *Proc Natl Acad Sci U S A*, 97(7):3154–9, 2000.

[95] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong, M. Shales, S. Lovett, M. E. Winkler, N. J. Krogan, A. Typas, and C. A. Gross. Phenotypic landscape of a bacterial cell. *Cell*, 144(1):143–56, 2011.

[96] T. UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 46(5):2699, 2018.

[97] A. Haldimann and B. L. Wanner. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *J Bacteriol*, 183(21):6384–93, 2001.

[98] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2:2006.0008, 2006.

[99] M. Kitagawa, T. Ara, M. Arifuzzaman, T. Ioka-Nakamichi, E. Inamoto, H. Toyonaga, and H. Mori. Complete set of orf clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 orf archive): unique resources for biological research. *DNA Res*, 12(5):291–9, 2005.

[100] Y. Ohashi, A. Hirayama, T. Ishikawa, S. Nakamura, K. Shimizu, Y. Ueno, M. Tomita, and T. Soga. Depiction of metabolome changes in histidine-starved *Escherichia coli* by CE-TOFMS. *Mol Biosyst*, 4(2):135–47, 2008.

[101] A. Hirayama, K. Kami, M. Sugimoto, M. Sugawara, N. Toki, H. Onozuka, T. Kinoshita, N. Saito, A. Ochiai, M. Tomita, H. Esumi, and T. Soga. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res*, 69(11):4918–25, 2009.

[102] R. Baran, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert, and M. Tomita. Mathdamp: a package for differential analysis of metabolite profiles. *BMC Bioinformatics*, 7:530, 2006.

[103] L. I. Hu, B. P. Lima, and A. J. Wolfe. Bacterial protein acetylation: the dawning of a new age. *Mol Microbiol*, 77(1):15–21, 2010.

[104] E. Masse and S. Gottesman. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 99(7):4620–5, 2002.

[105] I. Y. Shi, J. Stansbury, and A. Kuzminov. A defect in the acetyl coenzyme A<–>acetate pathway poisons recombinational repair-deficient mutants of *Escherichia coli*. *J Bacteriol*, 187(4):1266–75, 2005.

[106] C. Beloin, J. Valle, P. Latour-Lambert, P. Faure, M. Kzreminski, D. Balestrino, J. A. Haagensen, S. Molin, G. Prensier, B. Arbeille, and J. M. Ghigo. Global impact of mature biofilm lifestyle on *Escherichia coli* K-12 gene expression. *Mol Microbiol*, 51(3):659–74, 2004.

[107] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong, M. Shales, S. Lovett, M. E. Winkler, N. J. Krogan, A. Typas, and C. A. Gross. Phenotypic landscape of a bacterial cell. *Cell*, 144(1):143–56, 2011.

[108] R. V. Krishna, P. R. Krishnaswamy, and D. R. Rao. Enzymic synthesis of N-acetyl-L-phenylalanine in *Escherichia coli* K12. *Biochem J*, 124(5):905–13, 1971.

[109] D. C. Sévin, T. Fuhrer, N. Zamboni, and U. Sauer. Nontargeted *in vitro* metabolomics for high-throughput identification of novel enzymes in *Escherichia coli*. *Nat Methods*, 14(2):187–194, 2016.

Appendix A

Metabolome concentration of *yhhY* overexpressing and knockout cells

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values).

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
1-Methyl-2-pyrrolidinone	59.9	28.9	43.6	25.9	41.0	24.0
1-Methylnicotinamide	0.370	0.154	0.150	0.246	0.410	0.325
2-Amino-2-methyl-1,3-propanediol	1.50	1.61	1.21	1.71	2.49	1.52
2-Deoxyglucose 6-phosphate	0.497	0.707	0.894	1.02	1.47	0.932
2-Deoxyribose 1-phosphate	0.508	0.494	1.98	1.47	2.27	1.42
2-Furoate	8.04	9.27	18.3	11.2	13.8	9.86
2-Hydroxyglutarate	0.144	0.165	0.193	0.0609	0.0977	N.D.
2-Hydroxypentanoate	79.5	50.1	74.1	44.0	79.0	49.8
2-Oxoglutarate	1.98	2.40	5.87	3.08	N.D.	N.D.
2-Oxoisopentanoate	N.D.	1.43	2.84	1.52	1.80	N.D.
2,3-DPG	0.0657	2.10	0.129	1.76	3.14	N.D.
2'-Deoxyguanosine	0.411	0.359	0.621	1.71	2.03	1.74
2',3'-cCMP	0.122	0.182	0.336	0.215	0.150	N.D.

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
2AB	N.D.	0.362	0.605	0.388	0.666	0.452
2PG	2.41	4.46	33.6	30.7	62.6	31.1
3-Acetylacrylate	94.1	78.8	94.3	63.7	105	159
3-Hydroxybutyrate	4.67	3.38	4.63	2.40	4.93	3.47
3-Methylbutanoate	19.6	19.6	22.3	17.1	28.3	16.5
3-Phenylpropionate	9.57	7.09	N.D.	N.D.	N.D.	N.D.
3PG	13.0	26.9	27.2	31.1	63.5	30.3
4-Acetylbutyrate	32.3	23.1	29.7	17.7	34.5	19.1
4-Aminophenylsulfone	0.417	0.846	0.697	1.09	1.36	1.03
4-Methyl-2-oxopentanoate	2.26	2.72	3.83	1.91	3.46	2.08
4-Oxopentanoate	9.49	6.87	10.7	5.45	9.04	6.20
5-Hydroxyindoleacetate	N.D.	N.D.	N.D.	8.30	11.5	4.04
5-Methoxyindoleacetate	0.924	0.290	1.23	0.299	1.55	0.803
5-Methylthioadenosine	1.10	1.40	1.57	3.78	4.32	3.56
5-Oxoproline	6.61	3.41	6.61	6.24	7.33	8.15
5-Oxoproline	16.7	10.2	20.8	17.0	22.4	19.0
6-Hydroxyhexanoate	15.2	13.5	17.1	10.3	17.9	N.D.
6-Phosphogluconate	0.260	0.600	0.534	0.625	0.716	N.D.
Acetanilide	17.4	8.76	14.5	8.20	14.7	11.1
Acetyl CoA	486	260	254	1060	490	341
Adenine	2.17	2.26	3.01	5.07	8.07	6.32
Adenosine	0.422	0.360	0.608	1.70	1.99	1.72
Adenosine 3',5'-diphosphate	0.113	0.0711	0.181	0.119	0.114	N.D.
Adenylosuccinate	6.80	18.3	30.2	13.1	5.00	N.D.
Adipate	2.44	1.76	3.89	1.21	2.70	2.73
ADP	24.3	34.3	56.7	85.7	102	81.9
ADP-glucose	56.9	107	101	54.5	54.1	57.3
ADP-ribose	6.32	16.5	24.6	16.4	34.6	41.5
Ala	11.6	9.41	15.2	23.7	30.7	20.0
Ala-Ala	1.73	4.63	3.42	4.40	4.91	3.56

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Allantoin	2.27	9.31	3.89	6.81	9.06	5.86
α-Lipoamide	0.298	0.196	0.106	0.165	1.32	0.685
α-Methylserine	4.66	4.31	4.75	2.66	4.53	2.48
AMP	46.1	60.4	86.3	122	125	101
Anthranilate	0.312	0.184	0.230	0.347	0.502	0.406
Arg	12.8	12.4	11.4	15.2	11.6	14.8
Asn	1.26	1.34	1.04	0.988	1.66	1.06
Asp	4.67	5.77	5.63	8.02	14.6	7.55
Asp	8.27	13.1	12.2	17.0	29.8	12.5
ATP	25.3	51.5	60.5	72.9	58.0	60.3
Azelate	2.07	1.45	2.55	1.19	1.97	1.69
Benzamide	2.63	0.994	1.77	1.30	1.61	1.09
Benzoate	10.5	7.62	11.8	5.31	10.2	5.75
β-Leucine	9.40	8.03	12.6	13.5	17.2	9.38
Betaine	25.8	44.7	32.6	42.6	58.2	30.5
Butanoate	35.9	36.6	31.6	36.0	60.3	24.5
Cadaverine	0.705	0.232	0.458	1.31	1.58	1.43
cAMP	0.273	0.0788	N.D.	1.15	0.614	0.654
Carbachol	45.5	43.3	50.2	58.7	73.2	54.8
Carnitine	0.230	0.0831	0.164	0.0900	0.113	0.0630
CDP	28.7	23.1	37.0	26.2	31.8	25.6
cGMP	0.367	0.475	N.D.	1.30	1.16	0.818
Choline	8.11	12.5	12.8	19.6	34.6	20.8
cis-Aconitate	0.195	0.268	0.288	0.127	0.273	0.173
Citraconate	0.210	0.127	0.165	0.0677	0.139	N.D.
Citrulline	0.279	0.164	0.274	0.245	0.379	0.266
CMP	11.7	14.9	14.7	13.8	14.6	14.4
CoA	0.285	0.325	5.44	29.8	15.0	63.5
Creatine	1.22	0.489	0.787	0.504	0.776	0.669
Creatinine	0.397	0.193	0.365	0.210	0.358	0.349

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Crotonate	85.5	53.9	94.6	48.3	71.1	43.2
CTP	30.3	30.4	46.2	26.8	28.3	27.2
Cyclohexylamine	1.05	0.27	0.141	0.426	0.194	0.0781
dADP	9.31	6.94	14.2	11.5	14.7	11.0
dAMP	5.17	3.66	3.74	2.88	2.95	3.29
dATP	6.16	4.25	6.79	4.63	5.50	5.48
dCDP	3.80	3.92	5.28	5.69	5.68	5.85
dCMP	0.871	0.948	1.02	0.825	1.18	N.D.
dCTP	1.73	2.65	3.99	4.29	4.55	4.34
Deamido-NAD+	N.D.	2.19	3.59	3.55	3.58	5.11
Decanoate	7.87	16.0	22.9	14.4	21.3	4.45
dGTP	31.1	2.11	74.5	93.9	107	114
DHAP	7.49	10.9	9.04	17.2	12.4	12.2
Diethanolamine	1.41	1.51	1.13	1.61	2.34	1.43
Diethyl-2-phenylacetamide	0.0316	0.0134	0.0169	0.0169	0.0231	0.0130
Dihydrouracil	808	486	723	411	724	523
Dodecanedioate	0.207	0.206	0.228	0.157	0.235	0.200
dTDP	17.3	26.2	37.8	21.3	31.9	35.2
dTMP	5.80	9.79	14.4	5.65	9.20	11.3
dTTP	6.30	5.53	10.7	5.99	6.84	5.96
F1,6P	2.56	6.60	6.24	11.6	4.67	N.D.
F6P	0.600	1.72	4.62	4.05	6.59	3.96
FAD	2.13	2.12	3.00	5.04	4.40	4.23
Fumarate	4.24	6.21	5.31	5.19	5.69	4.80
G1P	1.41	4.79	3.93	3.31	3.54	N.D.
G3P	160	12.9	20.6	15.4	N.D.	15.4
G6P	784	2520	1650	1670	2720	1640
GABA	0.428	0.400	1.82	0.321	0.487	0.324
γ-Glu-2AB	0.195	0.168	0.505	0.641	0.883	0.648
GDP	33.6	43.7	56.7	88.6	99.8	85.0

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Gln	3.05	2.95	2.61	5.81	6.27	3.77
Glu	16.9	20.9	22.9	29.7	43.4	22.9
Glu	40.7	58.2	64.5	75.9	108	56.6
Glu-Glu	N.D.	0.275	N.D.	0.351	0.325	0.209
Glucosamine	0.210	0.286	0.249	0.367	0.547	0.288
Glucosamine 6-phosphate	4.49	11.3	6.90	13.3	13.6	N.D.
Glutarate	0.720	0.449	0.906	0.242	0.837	0.465
GSH	1.59	2.73	0.892	1.60	2.35	0.514
GSSG	6.03	11.7	3.88	6.59	9.62	2.39
Gly	7.94	7.26	8.60	7.65	10.7	9.00
Gly-Leu	0.184	0.272	0.185	0.266	0.218	0.181
Glycerophosphate	37.3	19.2	73.1	24.7	31.1	23.1
Glycolate	24.4	20.0	35.3	16.7	31.6	N.D.
GMP	12.6	14.4	17.3	27.8	24.9	33.0
Gramine	6.12	5.92	5.46	7.28	5.57	7.09
GTP	26.9	48.4	56.0	75.6	81.7	71.6
Guanine	1.29	0.923	1.42	0.985	1.59	1.46
Guanosine	2.13	1.13	1.53	1.53	1.37	2.08
Heptanoate	3.94	2.05	13.6	9.61	16.1	9.70
Hexanoate	15.5	11.2	31.2	20.8	35.8	23.7
His	4.19	4.52	5.73	4.22	4.61	2.93
Homoserine	1.56	1.53	1.79	4.05	4.73	3.84
Hypoxanthine	7.96	5.46	6.98	10.3	15.1	13.2
Ile	9.32	3.75	5.01	6.10	8.65	4.40
IMP	0.391	3.68	2.19	2.63	1.13	N.D.
Indole-3-acetamide	5.12	3.13	0.550	2.52	4.57	3.52
Indole-3-acetate	0.369	0.635	0.629	0.414	0.714	0.782
Indole-3-ethanol	1.73	1.56	0.404	0.628	0.541	0.689
Inosine	4.12	2.11	2.60	3.38	2.68	4.51
Isobutyryl CoA	0.387	0.374	0.709	0.505	1.31	1.05

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Isonicotinamide	4.13	9.54	14.5	9.49	22.6	12.2
Leu	9.54	8.15	12.8	13.7	17.4	9.52
Leu-Leu-Tyr	N.D.	N.D.	N.D.	1.40	1.66	1.11
Lys	105	100	116	136	170	127
Malate	10.7	18.1	16.9	14.8	17.8	15.2
Malonate	1.70	1.59	1.96	0.958	1.60	1.35
Melatonin	0.167	0.619	0.569	0.448	N.D.	N.D.
Met	1.75	2.24	2.67	2.97	3.65	2.00
Methyl sulfate	0.329	0.251	0.555	0.215	0.435	0.228
N-Acetylaspartate	0.523	2.57	0.331	0.309	0.675	0.228
N-Acetylglucosamine	1.14	0.490	0.510	0.366	0.794	0.572
N-Acetylglucosamine 1-phosphate	1.01	1.38	1.00	0.691	0.772	N.D.
N-Acetylglucosamine 6-phosphate	1.18	0.939	1.24	0.518	0.466	N.D.
N-Acetylglutamate	0.158	0.445	0.267	0.706	0.870	0.531
N-Acetylhistidine	N.D.	N.D.	N.D.	5.09	9.39	3.92
N-Acetylleucine	0.148	0.159	0.469	22.6	37.9	16.9
N-Acetylmethionine	N.D.	N.D.	N.D.	7.81	12.8	4.86
N-Acetylneuraminate	2.94	2.81	2.65	0.809	1.12	N.D.
N-Acetylornithine	N.D.	N.D.	N.D.	0.216	0.252	0.196
N-Acetylphenylalanine	N.D.	N.D.	N.D.	79.5	87.3	72.0
N-Acetylputrescine	7.07	1.55	4.89	1.38	2.54	1.56
N-Acetylvaline	N.D.	N.D.	N.D.	1.62	3.32	1.21
N-epsilon-Acetyllysine	0.380	0.561	0.381	1.54	2.55	1.38
N-Methylalanine	N.D.	0.343	0.574	0.368	0.632	0.428
N^6,N^6,N^6-Trimethyllysine	0.320	0.675	0.665	0.915	1.13	0.738
N^8-Acetylspermidine	0.194	0.0702	0.146	0.523	0.225	0.512
NADP+	152	129	179	242	235	224
Nicotinamide	3.87	8.94	13.6	8.90	21.2	11.4
Nicotinate	0.836	1.27	N.D.	1.40	N.D.	23.1
O-Acetylserine	18.1	22.5	24.7	32.0	46.7	0.312

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Octanoate	9.08	6.49	10.7	6.29	9.73	5.20
Octylamine	0.237	0.0771	0.0946	0.125	0.0834	0.0549
Ophthalmate	1.63	4.77	8.41	37.8	47.3	32.6
Ornithine	1.24	0.438	1.13	0.706	0.582	0.638
Pantothenate	0.711	1.63	0.694	0.665	0.839	N.D.
Pelargonate	8.42	5.09	10.9	12.4	11.9	12.3
Pentanoate	24.0	26.3	20.9	12.7	34.7	15.2
PEP	3.65	8.16	7.47	4.94	15.2	4.71
Phe	4.49	3.46	4.59	5.54	6.71	3.87
Phthalate	0.458	0.322	0.517	0.0589	0.336	0.202
Pimelate	0.724	0.428	0.628	0.324	0.594	0.742
Piperazine	188	91.8	151	95.0	145	99.9
Pro	3.41	6.20	4.62	6.09	9.65	4.69
Propionate	25.4	22.8	18.9	19.9	37.3	17.4
Putrescine	3.08	0.914	1.57	0.843	0.794	0.482
Pyridoxamine 5'-phosphate	0.393	0.797	0.658	1.03	1.29	0.973
R5P	1.97	3.49	3.29	3.95	3.97	3.13
Ribulose 1,5-diphosphate	0.264	0.408	0.388	0.232	0.592	N.D.
Ru5P	3.81	5.34	6.13	7.23	6.79	6.09
SAM+	1.16	0.682	0.726	3.06	2.34	1.47
Sarcosine	32.4	26.3	42.4	66.4	85.8	55.8
Sebacate	0.352	0.344	0.376	0.243	0.447	0.275
Ser	4.74	2.73	4.83	3.00	3.93	4.98
Sorbitol 6-phosphate	2.82	6.82	6.63	2.79	1.95	N.D.
Succinate	10.7	14.5	20.9	14.0	21.8	17.5
Taurine	3.70	1.56	2.79	1.68	1.94	1.97
TDP-glucose	51.5	47.2	50.3	29.5	30.9	27.5
Tetrahydropalmatine	N.D.	0.0227	0.0216	0.104	0.0822	0.0606
Thr	4.48	4.14	4.56	2.56	4.36	2.38
threo-beta-methylaspartate	40.1	57.4	62.8	74.5	107	55.9

Table A.1 Metabolome concentration of *yhhY* overexpressing cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3
Thymidine	10.9	4.29	6.62	3.68	5.41	6.22
Tiglate	7.05	6.68	6.19	5.76	13.1	N.D.
trans-4-Hydroxy-3-methoxycinnamate	2.29	1.44	2.57	1.10	2.21	1.60
Triethanolamine	1.66	0.312	0.804	0.389	N.D.	0.440
Trp	0.490	0.286	0.357	0.168	0.300	0.175
Tyr	1.17	1.58	1.50	2.36	3.19	1.73
Tyrosine methyl ester	20.6	22.6	27.0	23.9	28.0	22.9
UDP	21.7	25.3	33.8	36.0	48.7	33.4
UDP-glucose	43.7	69.7	54.7	100	84.2	75.3
UDP-glucuronate	3.63	3.83	3.24	2.26	2.09	2.39
UDP-*N*-acetylglucosamine	36.0	52.5	71.1	48.9	50.1	49.2
UMP	11.3	14.6	13.4	21.0	17.7	21.5
Undecanoate	1.55	1.40	3.69	1.93	3.37	2.12
Urea	26.8	17.8	27.4	15.4	22.7	28.8
Uridine	5.77	3.79	5.30	3.61	6.42	4.62
Urocanate	1.06	0.570	1.09	0.405	0.798	0.881
UTP	16.9	25.1	35.3	26.0	36.4	27.5
Val	6.18	5.89	7.92	10.2	14.3	7.21
XMP	0.301	1.37	0.732	0.824	0.552	0.830

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
2-Deoxyglucose 6-phosphate	0.730	1.05	N.D.	N.D.	N.D.	N.D.	3.02	3.04	11.1	17.2	N.D.
2-Deoxyribose 1-phosphate	1.96	1.43	2.62	2.41	1.31	2.06	1.68	1.73	4.03	2.80	0.820
2-Furoate	8.97	10.5	11.4	7.61	6.25	9.36	11.7	6.29	6.60	8.77	6.40
2-Hydroxy-4-methylpentanoate	1.31	2.13	N.D.	1.80	1.44	1.38	1.37	2.47	N.D.	N.D.	N.D.
2-Hydroxybutyrate	1.93	3.06	2.06	3.41	N.D.	N.D.	N.D.	1.92	1.26	2.41	4.75
2-Hydroxyglutarate	49.8	63.9	102	74.3	201	356	69.0	127	N.D.	537	N.D.
2-Hydroxyisobutyrate	1.54	2.45	1.65	2.73	N.D.	N.D.	N.D.	1.54	2.69	1.93	3.74
2-Hydroxyoctanoate	N.D.	N.D.	0.460	0.310	N.D.	N.D.	0.370	0.450	N.D.	0.570	N.D.
2-Hydroxypentanoate	104	117	110	98.6	125	119	101	129	98.9	135	112
2-Isopropylmalate	7.35	8.12	8.53	19.0	18.3	28.0	19.9	30.6	N.D.	157	N.D.
2-Oxoglutarate	48.2	65.0	76.0	64.5	20.1	34.5	50.2	58.8	373	301	N.D.
2-Oxoisopentanoate	9.72	10.7	13.8	14.4	N.D.	8.78	19.6	27.6	94.0	87.7	13.4
2,4-Dihydroxypyrimidine-5-carboxylate	121	41.2	65.4	71.8	54.1	32.6	51.1	50.2	223	161	227
2AB	7.99	3.72	11.8	1.76	3.51	5.72	15.2	7.91	17.3	8.60	42.4
3-Hydroxybutyrate	5.93	5.93	5.67	7.21	7.79	6.71	5.76	6.06	4.45	11.6	7.21
3-Hydroxypropionate	N.D.	N.D.	16.8	17.3	11.0	13.2	N.D.	N.D.	550	15.5	101
3-Methylbutanoate	32.1	43.1	39.5	38.0	29.9	37.2	41.4	46.1	44.3	74.2	N.D.
3-Phenylpropionate	11.6	11.2	15.0	11.3	N.D.	12.6	10.6	11.3	16.2	N.D.	N.D.
3PG	50.6	73.3	216	132	130	80.5	29.9	72.9	N.D.	342	N.D.
4-Acetylbutyrate	47.4	48.0	47.2	41.4	53.0	45.9	45.1	50.6	43.6	54.3	38.0
4-Methyl-2-oxopentanoate	6.86	7.64	6.68	5.17	5.11	3.65	10.5	20.1	64.3	44.7	N.D.
4-Methylbenzoate	262	252	250	221	287	248	269	288	290	376	278
4-Oxohexanoate	1.22	1.57	0.840	0.970	53.3	0.820	0.880	1.02	43.8	1.93	38.2
4-Oxopentanoate	13.6	15.7	15.2	15.4	18.2	16.6	17.4	14.3	14.6	20.9	13.3
5-Methylthioadenosine	5.47	4.77	7.46	7.62	3.57	2.00	7.20	6.20	23.8	10.4	3.02
5-Oxoproline	20.0	32.9	67.3	63.5	29.6	43.4	34.8	45.0	57.4	75.7	46.2
6-Hydroxyhexanoate	34.7	31.6	43.0	41.3	25.4	30.8	33.9	38.1	19.8	22.0	N.D.
6-Phosphogluconate	8.25	4.17	8.53	14.2	21.9	15.8	2.27	2.75	12.3	14.6	2.51

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
Acetyl CoA	39.6	168	N.D.	2.04	N.D.	0.560	77.1	165	44.7	N.D.	N.D.
Adenine	10.1	11.6	17.7	18.4	8.75	5.35	12.0	11.5	36.4	66.7	1.64
Adenosine	2.70	3.87	3.86	5.20	1.45	1.35	5.64	7.37	84.2	440	1.86
Adenosine 3',5'-diphosphate	3.15	N.D.	N.D.	5.06	N.D.	N.D.	N.D.	11.1	23.3	25.0	N.D.
Adenylosuccinate	0.920	0.960	1.19	1.37	1.46	0.570	N.D.	1.03	19.5	3.05	N.D.
Adipate	4.07	3.95	5.64	5.37	3.39	3.28	8.00	5.02	1.86	2.94	N.D.
ADP	132	130	173	228	187	141	103	79.5	549	254	7.23
ADP-glucose	2.53	1.39	N.D.	N.D.	2.20	N.D.	N.D.	N.D.	1.48	N.D.	N.D.
ADP-ribose	14.5	5.95	18.2	5.19	3.56	6.05	6.85	4.49	0.440	46.6	N.D.
Agmatine	N.D.	N.D.	0.400	0.550	0.260	0.350	1.36	0.720	0.990	0.500	3.46
Ala	43.9	62.7	117	89.9	89.3	142	57.0	62.2	289	159	231
Ala-Ala	8.86	11.1	18.4	23.8	24.2	27.4	13.0	14.9	233	67.7	22.5
α-Aminoadipate	N.D.	N.D.	N.D.	N.D.	1.27	0.690	N.D.	N.D.	2.79	N.D.	1.36
α-Methylserine	36.1	38.6	N.D.	67.7	453	518	36.8	38.8	12.5	3.30	83.5
AMP	41.6	47.8	46.6	61.8	72.8	45.4	43.2	29.9	326	132	N.D.
Arg	28.8	36.2	63.3	33.6	42.6	33.3	32.0	29.5	81.4	62.9	25.2
Asn	12.6	16.6	32.4	24.0	28.1	33.1	19.2	18.3	111	57.9	68.7
Asp	10.9	13.5	43.1	36.5	28.1	31.0	17.3	19.2	60.4	50.2	261
Asp	12.9	18.2	40.4	39.2	32.1	36.7	22.4	25.9	99.4	50.8	256
ATP	451	441	649	669	595	487	237	313	898	805	9.98
Azelate	2.29	1.98	2.73	3.16	1.92	2.21	2.44	2.35	1.70	2.10	1.14
Benzoate	10.6	16.2	14.1	15.0	N.D.	18.3	14.8	15.8	14.7	N.D.	26.1
β-Ala	1.93	2.08	2.17	2.18	1.34	0.980	2.02	1.64	3.49	4.36	N.D.
β-Ala-Lys	N.D.	2.16	3.75	2.31	1.81	2.26	N.D.	N.D.	2.52	1.72	N.D.
Betaine	4.91	4.34	20.6	4.88	3.28	3.92	5.35	3.54	4.02	N.D.	6.00
Butanoate	46.5	80.2	68.2	60.8	69.1	83.8	71.0	91.4	87.4	123	40.5
Cadaverine	N.D.	N.D.	1.31	0.580	2.96	0.580	1.19	0.940	0.350	1.58	3.79
cAMP	0.650	2.01	9.25	N.D.	0.600	N.D.	0.900	2.15	0.220	0.540	N.D.

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
Carbamoylaspartate	107	104	121	131	72.7	54.0	72.1	47.8	698	458	78.3
CDP	30.5	35.7	37.1	33.4	52.0	39.7	31.3	17.3	110	36.7	2.08
Choline	1.25	1.08	2.04	0.960	1.04	1.36	2.91	3.52	0.870	2.26	N.D.
cis-Aconitate	3.27	2.80	4.60	3.94	5.85	5.80	3.24	4.40	32.2	16.9	N.D.
Citraconate	N.D.	0.780	0.930	N.D.	0.790	0.990	N.D.	N.D.	6.46	1.71	N.D.
Citrate	75.5	71.5	130	106	189	170	50.9	106	N.D.	N.D.	37.5
Citrulline	7.87	9.12	14.0	18.3	9.63	11.1	12.7	13.6	76.2	58.6	14.3
CMP	23.9	26.6	30.3	19.1	37.2	30.6	27.3	24.8	90.3	55.3	1.22
CMP-*N*-acetylneuraminate	24.8	28.0	59.6	59.3	43.5	57.8	27.0	24.8	176	89.6	8.29
CoA	N.D.	0.160	N.D.	N.D.	6.14	6.47	N.D.	N.D.	1.17	0.280	N.D.
CTP	107	117	179	121	158	143	60.0	71.3	155	135	8.35
Cytidine	N.D.	0.780	0.660	1.03	N.D.	N.D.	1.21	1.60	11.9	79.4	N.D.
Cytosine	0.980	1.09	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	1.23	1.25	N.D.
dADP	12.8	12.9	16.9	25.4	21.1	15.6	7.41	5.97	47.4	23.2	N.D.
dAMP	1.95	2.85	4.06	3.48	3.90	2.16	1.32	1.26	11.5	4.61	N.D.
dATP	30.2	30.6	48.8	53.2	42.3	33.8	11.2	13.9	51.1	52.2	0.470
dCDP	4.70	5.57	6.87	5.90	6.62	4.75	3.76	2.21	24.3	9.87	N.D.
dCMP	0.770	0.920	1.81	0.890	0.860	N.D.	0.700	0.780	4.42	2.18	N.D.
dCTP	15.1	15.3	23.1	16.8	17.5	16.2	8.33	8.55	17.6	19.2	1.76
Decanoate	30.3	22.0	25.6	23.9	21.0	22.5	25.8	22.2	N.D.	3.62	N.D.
dGTP	14.2	19.9	19.4	26.8	9.43	10.1	5.08	7.61	435	9.98	N.D.
DHAP	22.0	18.1	14.8	20.8	35.1	32.3	23.1	24.1	28.6	38.4	59.3
Diethanolamine	11.5	20.7	9.99	14.7	27.0	8.54	17.7	15.0	14.5	15.1	12.7
Dihydroorotate	19.0	22.6	34.6	37.9	19.6	15.5	23.5	15.8	370	177	92.6
dITP	N.D.	N.D.	N.D.	N.D.	N.D.	4.62	N.D.	3.28	4.90	7.95	N.D.
Dodecanedioate	0.390	0.340	0.400	0.300	N.D.	0.380	0.240	0.380	N.D.	0.340	N.D.
DOPA	N.D.	3.52	5.89	6.07	3.65	3.24	2.67	1.73	13.1	9.52	2.68
dTDP	12.5	13.8	19.9	16.8	22.2	17.5	8.99	6.20	45.8	19.8	N.D.

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
dTMP	7.19	5.15	25.3	12.2	9.81	11.6	8.59	4.14	69.4	165	N.D.
dTTP	36.0	36.3	63.6	37.0	52.0	46.9	17.0	19.0	50.5	42.4	2.12
dUDP	0.530	0.400	0.390	0.610	0.640	0.260	0.490	0.280	1.59	1.37	N.D.
dUMP	0.940	1.21	0.22	N.D.	N.D.	N.D.	1.73	5.21	3.20	0.800	N.D.
dUTP	2.28	2.02	2.55	1.43	1.71	1.33	1.24	1.03	1.47	1.36	N.D.
F1,6P	59.5	58.6	55.5	102	97.3	77.2	7.15	14.1	N.D.	4.57	N.D.
Fumarate	26.4	24.8	46.0	49.5	23.3	16.5	41.7	42.1	N.D.	142	N.D.
G1P	4.57	4.81	3.86	N.D.	N.D.	0.180	11.0	11.2	28.4	90.4	2.85
GABA	5.20	6.74	15.3	14.2	16.4	20.4	6.93	6.12	56.7	48.3	4.98
γ−Glu-2AB	1.88	2.57	3.94	1.21	3.48	3.24	8.05	4.20	5.67	4.13	9.66
γ−Guanidinobutyrate	N.D.	0.390	N.D.	N.D.	0.290	N.D.	0.480	0.650	0.990	1.00	N.D.
GDP	52.6	61.0	62.2	90.0	76.3	56.6	41.3	33.7	277	76.0	N.D.
Gln	12.3	16.2	62.0	92.4	47.9	99.2	39.1	36.3	300	194	375
Glu	589	754	1690	2100	817	1500	834	735	5550	4130	4610
Glu-Glu	0.910	N.D.	2.23	2.01	1.43	1.80	N.D.	N.D.	1.29	N.D.	N.D.
Glucosamine	1.25	1.36	2.39	1.68	1.62	1.39	3.43	3.23	3.70	4.42	1.89
GSH	4.81	7.12	N.D.	2.27	26.0	34.2	2.23	N.D.	1.44	3.76	N.D.
GSSG	231	288	597	628	485	600	310	301	1550	1130	499
Gly	50.3	65.4	124	82.0	73.2	79.6	78.2	148	286	194	147
Glycerate	8.64	6.17	10.2	18.8	6.21	8.29	N.D.	6.54	21.8	29.9	N.D.
Glycerophosphate	13.0	13.9	19.5	20.3	24.2	23.4	13.5	11.6	55.3	29.8	10.6
GTP	142	160	177	236	211	155	63.7	84.0	325	272	N.D.
Guanine	N.D.	4.40	8.12	3.51	5.35	2.27	2.35	5.12	1.67	4.42	N.D.
Guanosine	5.55	7.77	13.1	6.05	4.00	3.10	4.88	6.60	55.3	140	N.D.
Heptanoate	24.4	23.6	24.8	20.5	22.9	21.8	23.2	24.4	23.5	39.5	N.D.
Hexanoate	21.2	17.5	21.0	19.0	N.D.	16.9	19.7	N.D.	50.1	35.5	N.D.
His	7.59	11.4	22.6	15.3	15.2	13.0	5.63	8.42	12.1	11.8	7.81
Histidinol	0.180	N.D.	0.830	0.620	1.3.0	0.600	N.D.	N.D.	0.260	N.D.	N.D.

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
Homoserine	8.13	7.81	12.2	11.8	27.9	31.1	11.0	12.2	26.2	19.3	28.0
Hypoxanthine	15.3	22.7	55.7	25.6	25.3	9.68	12.1	20.6	7.45	27.1	N.D.
Ile	2.58	2.82	6.80	5.06	8.88	11.5	4.28	45.9	110	16.6	10.2
Inosine	12.4	3.43	27.2	11.1	8.03	5.25	N.D.	4.73	5.84	18.5	N.D.
Isobutyryl CoA	1.14	2.68	0.430	1.80	N.D.	N.D.	0.770	2.85	0.240	N.D.	N.D.
Isocitrate	160	N.D.	N.D.	0.870	405	N.D.	0.670	225	N.D.	N.D.	225
Isonicotinamide	39.0	15.0	67.5	45.6	16.1	19.9	19.0	14.0	122	218	5.15
Itaconate	N.D.	0.610	N.D.	0.770	0.730	0.550	0.780	N.D.	N.D.	N.D.	N.D.
ITP	48.6	46.4	70.1	69.3	57.6	50.8	25.9	34.7	83.0	85.9	1.19
Lactate	404	326	364	377	364	355	446	739	366	605	698
Leu	11.4	14.7	20.0	27.0	19.2	16.3	23.0	46.2	111	84.4	35.0
Lys	104	134	226	141	88.6	109	155	187	490	394	N.D.
Malate	48.5	48.0	77.9	78.9	66.4	50.4	62.6	84.3	N.D.	235	33.2
Malonate	3.21	3.91	5.11	4.03	4.12	3.83	3.99	3.29	N.D.	8.38	N.D.
Mannosamine	N.D.	N.D.	N.D.	N.D.	1.58	N.D.	N.D.	N.D.	1.20	4.22	3.34
Met	9.05	11.4	23.7	11.7	7.12	6.16	28.2	44.3	140	62.1	19.6
Methanesulfonate	2.10	3.95	3.05	4.44	2.96	3.57	4.21	4.23	N.D.	N.D.	N.D.
N-Acetylaspartate	43.8	45.3	42.3	128	48.8	25.0	60.8	62.9	N.D.	344	N.D.
N-Acetylglucosamine 1-phosphate	N.D.	1.74	2.27	N.D.	0.480	N.D.	N.D.	N.D.	4.23	19.5	N.D.
N-Acetylglucosamine 6-phosphate	1.59	2.00	2.34	2.34	1.34	1.24	0.770	0.640	1.46	1.13	N.D.
N-Acetylglutamate	0.530	0.900	2.69	13.7	6.59	3.37	2.77	1.83	79.2	29.5	18.3
N-Acetylmethionine	0.670	0.500	N.D.	0.440	N.D.	N.D.	0.320	N.D.	0.410	N.D.	N.D.
N-Acetylmuramate	0.810	0.810	1.14	1.30	N.D.	N.D.	0.530	1.23	N.D.	N.D.	N.D.
N-Acetylornithine	3.17	4.28	16.8	23.3	20.6	16.4	2.84	N.D.	36.0	17.8	6.53
N-Acetylputrescine	1.37	1.20	4.83	3.07	1.72	1.59	2.12	2.46	6.77	6.16	3.07
N^6,N^6,N^6-Trimethyllysine	N.D.	N.D.	0.700	N.D.	0.510	0.400	N.D.	0.680	1.25	1.17	N.D.
N^8-Acetylspermidine	0.290	0.640	0.920	0.890	0.310	0.420	1.09	0.610	2.19	0.920	1.10
NAD+	N.D.	N.D.	271	198	211	172	130	140	N.D.	217	N.D.

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
NADPH	1.42	2.30	0.820	3.33	N.D.	0.810	0.830	1.66	2.53	8.55	0.970
Nicotinamide	31.4	12.0	54.2	36.6	13.0	16.0	15.3	11.2	98.3	175	4.14
o-Hydroxybenzoate	N.D.	1.63	1.36	N.D.	N.D.	0.490	N.D.	1.56	N.D.	N.D.	N.D.
O-Phosphoserine	3.47	2.25	2.87	4.47	3.03	4.67	2.96	2.92	21.9	13.0	33.8
Octanoate	13.5	12.3	12.9	10.8	11.2	13.0	12.4	12.2	5.09	20.2	N.D.
Ophthalmate	87.9	77.7	119	25.4	96.3	146	187	116	269	179	159
Ornithine	4.39	11.0	20.6	11.5	7.84	7.83	9.82	11.8	37.5	19.4	13.6
P1, P4-Di(adenosine-5') tetraphosphate	1.44	1.75	3.09	2.55	N.D.	2.59	1.26	1.10	0.320	2.07	N.D.
Pantothenate	N.D.	N.D.	2.92	2.98	N.D.	N.D.	0.200	N.D.	N.D.	18.3	N.D.
Pelargonate	35.0	30.2	33.0	29.6	32.7	32.0	31.5	32.6	32.6	55.0	N.D.
PEP	15.4	26.0	82.2	38.0	31.4	21.5	19.1	29.2	N.D.	N.D.	0.940
Phe	9.35	9.47	17.4	16.7	12.9	11.9	9.68	11.8	47.7	39.3	15.9
Phenylpyruvate	1.29	1.67	1.48	N.D.	N.D.	N.D.	N.D.	N.D.	2.30	0.510	N.D.
Phthalate	0.760	0.770	1.09	0.840	0.720	0.750	0.850	1.09	0.260	0.510	N.D.
Pimelate	1.58	1.92	1.93	1.53	1.28	1.52	2.08	1.61	N.D.	2.79	N.D.
Pro	16.7	13.8	33.8	26.2	21.0	16.2	15.0	20.0	52.3	45.9	34.9
Propionate	18.2	31.0	28.6	25.3	26.6	31.0	23.4	34.2	32.2	58.2	N.D.
Pterin	9.61	12.0	20.7	22.4	13.9	14.7	23.0	38.3	174	68.8	25.2
Putrescine	121	175	269	196	265	213	526	358	881	610	N.D.
Pyridoxamine 5'-phosphate	3.78	4.35	7.30	8.46	8.95	6.40	3.62	5.09	25.5	10.8	10.0
R5P	19.4	16.0	14.7	N.D.	N.D.	N.D.	15.8	26.1	59.9	N.D.	4.17
Ribulose 1,5-diphosphate	3.38	5.62	9.06	6.49	8.50	6.23	1.13	3.54	N.D.	5.44	N.D.
Ru5P	40.3	30.9	41.8	35.8	45.8	65.3	23.9	23.6	63.5	61.8	19.7
S7P	11.7	9.69	13.3	14.2	N.D.	22.7	12.8	13.1	62.0	52.9	2.50
SAM+	5.53	8.51	8.03	9.24	6.40	3.36	10.5	10.3	29.7	10.7	16.6
Sebacate	0.490	0.460	0.630	0.530	0.480	0.580	0.540	0.590	0.660	0.540	0.310
Ser	26.0	45.9	87.5	57.4	21.9	37.3	35.0	44.4	110	95.8	39.8
Sorbitol 6-phosphate	11.8	10.3	12.6	27.1	23.4	17.3	13.4	14.9	52.6	79.1	4.21

Table A.2 Metabolome concentration of *yhhY* knockout cells (μM, N.D.: not detectable values, continued)

Metabolite name	Ctrl1	Ctrl2	Ctrl3	*yhhY*1	*yhhY*2	*yhhY*3	Ctrl+Co1	Ctrl+Co2	*yhhY*+Co1	*yhhY*+Co2	*yhhY*+Co3
Spermidine	1.19	0.700	1.40	1.67	1.10	0.750	1.60	1.01	N.D.	5.45	N.D.
Succinate	45.7	36.2	95.5	69.8	34.4	31.4	48.1	38.0	223	187	3.56
Succinyl CoA	N.D.	N.D.	N.D.	0.540	N.D.	0.540	N.D.	N.D.	0.410	2.24	N.D.
TDP-glucose	2.18	1.96	N.D.	6.34	4.19	2.45	3.15	N.D.	N.D.	N.D.	N.D.
Terephthalate	0.270	N.D.	N.D.	N.D.	1.14	0.270	0.150	0.140	N.D.	N.D.	N.D.
Thr	40.4	43.1	58.1	75.6	506	579	41.2	43.4	139	115	93.3
threo-beta-methylaspartate	447	593	1230	1470	669	1200	619	540	N.D.	3190	3800
Tiglate	20.7	21.2	25.2	21.2	20.4	18.3	19.2	23.5	28.0	45.1	N.D.
trans-Cinnamate	N.D.	27.0	50.9	67.8	25.1	41.2	24.3	21.2	N.D.	N.D.	N.D.
Trehalose 6-phosphate	N.D.	N.D.	N.D.	N.D.	N.D.	1.19	1.17	0.900	3.54	1.55	N.D.
Triethanolamine	N.D.	2.22	2.33	2.09	1.91	2.35	2.01	1.89	1.15	1.40	2.66
Trimethylamine *N*-oxide	0.870	1.11	1.04	1.09	1.10	1.32	1.64	1.56	1.48	1.01	1.83
Trp	2.37	2.35	4.85	3.14	3.34	3.51	2.07	1.97	9.48	5.70	2.87
Tyr	6.86	8.37	19.4	10.7	16.2	13.3	7.05	6.63	30.9	19.3	6.87
Tyramine	N.D.	N.D.	N.D.	N.D.	1.24	N.D.	1.79	0.950	1.46	N.D.	7.94
UDP	N.D.	N.D.	162	108	127	87.2	77.5	86.1	392	98.4	0.460
UDP-glucose	398	408	558	412	382	357	424	300	661	468	223
UDP-glucuronate	32.8	34.0	50.0	56.9	57.3	45.7	38.4	32.5	127	65.7	0.710
UDP-*N*-acetylglucosamine	161	156	223	136	114	N.D.	142	120	270	229	15.0
UMP	25.4	23.3	72.9	35.2	33.6	23.8	33.8	24.8	186	223	0.790
Undecanoate	2.60	2.14	1.67	2.72	N.D.	2.00	N.D.	N.D.	3.52	N.D.	N.D.
UTP	244	279	495	243	271	225	125	234	386	307	11.4
Val	31.2	37.1	87.3	35.3	49.2	69.6	56.6	73.6	287	178	152
XMP	1.64	1.74	1.96	2.59	1.06	0.770	2.73	2.90	4.61	4.78	N.D.

Acknowledgements

These works were supported by the Taikichiro Mori Memorial Research Fund and the research funds from Yamagata Prefectural government and the City of Tsuruoka.

I thank my sincere gratitude to Professor Masaru Tomita for the support from bachelor to doctor course. I express my deep appreciation to Project Professor Masahiro Sugimoto for his valuable comments. Dr. Martin Robert always kindly taught me scientific thinking and writing in the master's course. Associate Professors Yasuhiro Naito and Hiroki Kuroda carefully checked this thesis and gave me many useful comments. I express my sincere gratitude to Professors Mitsuhiro Itaya, Akio Kanai, and Tomoyoshi Soga. They always gave me valuable advice and led the direction of the study.